# Processing, Analysis and Visualization of Social Data

*Thesis submitted in partial fulfilment of the requirements for the*

*award of degree of*

**Master of Engineering**

**In**

**Computer Science and Engineering**


*Submitted By*

**Neha Garg**

**(801532033)**

Under the supervision of:

**Dr. Rinkle Rani**

Associate Professor



COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY
PATIALA – 147004


**JULY 2017**

# CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Processing, Analysis and Visualization of Social Data*", in partial fulfillment 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.

Neha Garg
**(Neha Garg)**

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

Rinkle

**(Dr. Rinkle Rani)**
Associate Professor
Computer Science and Engineering
Department

# 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**, Associate Professor, Computer Science and Engineering Department, Thapar University for her positive attitude, constant encouragement, keen interest, invaluable cooperation, generous attitude and above all her blessings. She has been a source of inspiration for me, I am grateful to **Dr. Maninder Singh**, Head of Department and **Dr. Ashutosh Mishra**, P.G. Coordinator, Computer Science and Engineering Deapartment, 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.

Later 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.

Neha Garg

Neha Garg
(801532033)

# ABSTRACT

Socializing is central to the nature of humans and it is widespread, but scientists have long pondered how it can be analyzed and explored. One answer involves examining the networked interactions and population structure available on social media like Facebook, Twitter, and LinkedIn etc. A social structure of individuals related directly or indirectly on the basis of some common factor like similar likings etc. is a social network. In order to understand the behavior and structure of a social network we need to study the network and this study is called social network analysis. There has been a rapid increment in the research and study of data mining community and social network analysis. There are various social networking sites available on internet like LinkedIn, Facebook, Instagram, Twitter, Google and many more. Interactions over such sites produces huge amount of data because billions of active users maintain their accounts. Hence, it is a tedious task to analyze the complex data. People often are in control over whom they interact with that comprises their social circle. Social media provides a platform to people to express their thoughts. The user is free to enter the text in any form. As a result there is a possibility of inconsistency. To remove these inconsistencies, first the data is normalized on the basis of some transformation. It is of great importance for academic and business to analyze such online social communities and predicting their behavior. In this research, LinkedIn data is extracted and apply the normalization technique to remove the redundancies. Then apply the Hierarchical Clustering algorithm to the normalized data set to cluster the data according to the job title and visualized the clustered data in the form of tree and dendrogram.

Secondly, we extract tweets about "MCDResults" (MCD results of Delhi) using Twitter API and R-tool for social data analysis. Then apply the preprocessing technique to clean the data for the further analysis and cluster the tweets based on the geolocation information using k-means clustering algorithm.

We also describe the different approaches in the field of community detection and compared those approaches based on the modularity function using different datasets of real world network of varying sizes.

# TABLE OF CONTENTS

# LIST OF FIGURES

| Figure No. | Description | Page No. |
|---|---|---|

# LIST OF TABLES

# Chapter 1 INTRODUCTION

Network is present everywhere on the web which is the most significant web network comprises of vertices represented by the billion of pages and edges represented by the hyperlink to each other [2]. The other form of networks such as query graph is formed by collecting and processing the inputs of web users [1]. The more networks are even created by using the Social Media applications such as YouTube, Facebook, IMDB, Twitter etc. Social Media Data which is originating from the networks is a very popular domain of research and here we shall refer this as a Social Media Network. Fig 1.1 shows the different aspects of social media of different social media companies.


**Fig. 1.1: Social Media Landscape**

A significant source of intelligence is presented by Social Media Networks to encode the online actives and source of information given by social networking members, in spite of their disparities as for the elements and the sort of connection they show. Because of the tremendous measure of information and very powerful nature, the major challenge for data mining methods to analyze such networks.

Today, the study area for data mining research is to collect the social media data and to analyze them to find the meaningful knowledge pattern and trends. Social media such as Orkut, Facebook, Instagram, LinkedIn, Twitter, Pinterest, Google+ etc. allows the individual to share their ideas, opinions in various fields and also allows the organization to advertise their products and services. The public opinion provides the platform to serve the feedback for those organizations. Some facts about the social media can be analyzed by Search Engine Journal and wrote that "93% of marketers are using Social Media" [3].

Community Detection is a significant tool for the analysis of substantial complex networks. The community detection in networks is the clustering of networks into communities in such a way that the connections between nodes within the same community are strong as compared to the connections between the nodes across the community. To detect and analyze the communities in a network become popular in domains ranging from social to biology and the Web. An extensive variety of intelligent services and applications such as automatic event detection, recommendation engines in Social Media content and so on can be exploited by detecting the communities on Social Media networks.

## 1.1 Social Media

Social Media provides the interactive platform that allows the user to create, share and exchange their thoughts and information through virtual groups and systems. Andreas M. Kaplan and Michael Haenlein who are the two professionals of Marketing and Customer Relations refer Social Media as "the services offered to users, Web 2.0 to the technologies that enable the easy use of the services and User Generated Content to the images, audio, videos, text etc. produced [4]". The pervasive and substantial changes are introduced by the Social Media to communicate between the individuals, communities and organization [5]. Social Media is different from traditional/industrial

media is various aspects such as frequency, quality, reach, permanence, usability and immediacy. According to Nielsen, the users spend more time with social media sites than any other site [6].

Many different forms are taken by social media technologies which includes social blogs, social networks, videos, microblogging, Internet forum, social bookmarking, weblogs, rating and podcasts. Technologies include: wall-posting, picture-sharing, vlogs, crowdsourcing, music-sharing and voice over IP. Many of the services can be integrated via social network aggregation platforms. Kaplan and Haenlein proposed the six different categories of social media services: content communities (for example, YouTube and DailyMotion), virtual game worlds (for example, World of Warcraft), collaborative projects (for example, Wikipedia), social networking sites (for example, Orkut, Facebook), virtual social worlds (for example, Blue Mars and Twinity) and blogs and microblogs (for example, Tumblr, Twitter).

## 1.2 Community Detection

In the real world, there are many networks and we focus on the social networks which involves the social interaction between humans. One such network is the *biological network* which shows the relationship between the molecular components. In this network each molecular component is represented by a node or vertex and their direct or indirect interaction is represented by an edge. In the real world networks, the vertices are group together into communities such that more edges within the communities rather than across the communities. The main focus of community detection is to reveal the hidden communities of a network. A graph having 34 nodes connected together is shown in Fig 1.2. The graph is divided into 4 different communities based on some common property. Social network is represented as a graph.

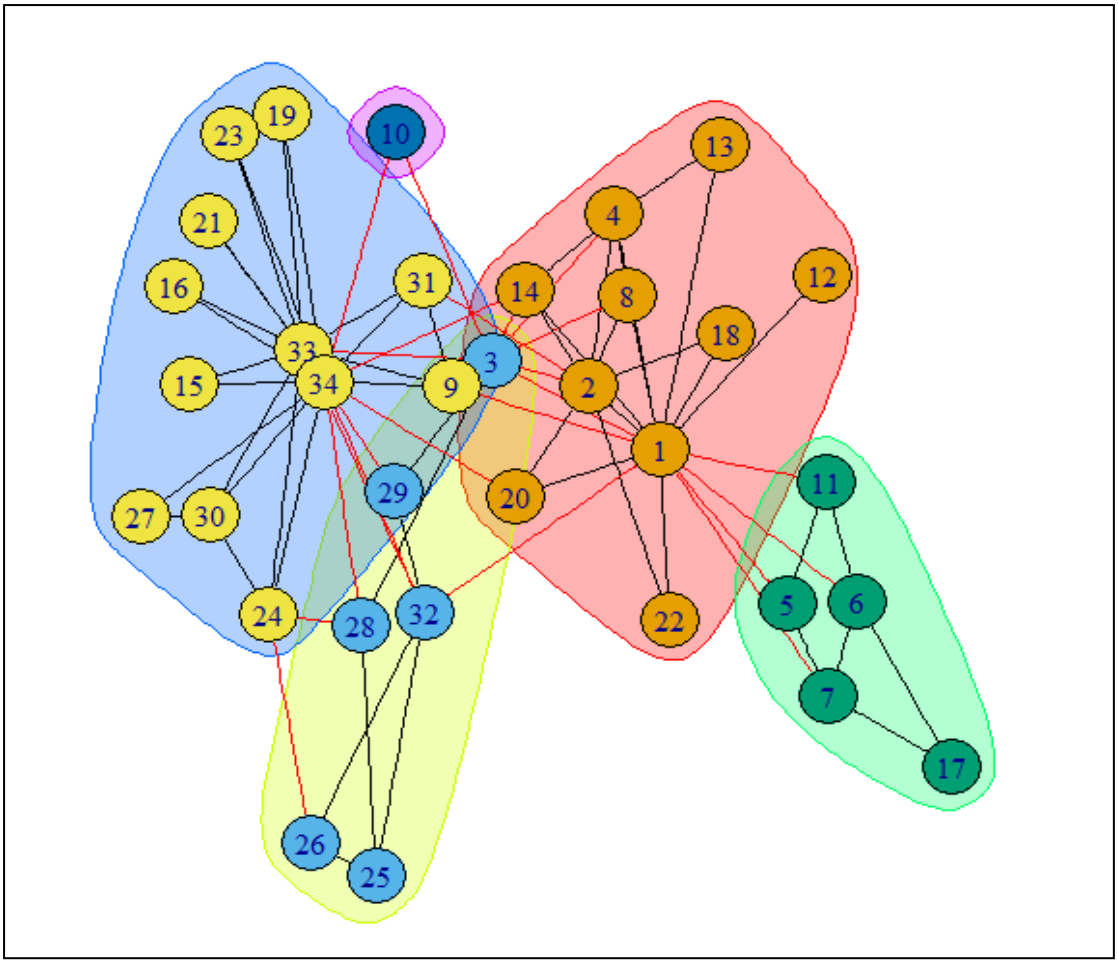


**Fig. 1.2: Community in Social Network**

There are various examples of real world networks including collaboration networks, social networks and computer networks.

Computer networks consist of computers represented by vertices which are connected by the connections such as satellite or cable and connection is represented by edges.

Computer networks consist of computers represented by vertices which are connected by the connections such as satellite or cable and connection is represented by edges. Various types of technological networks are exists including transportation networks and phone networks.

Virtual technological networks such as World Wide Web (WWW) consists of a directed network of HTML pages represented by vertices and connection between the HTML pages are represented by edges. Another type of network is the email network in which vertex represent the each individual and relationship between them is

represented by a directed edge. A directed edge (m, n) exists if n is in the address book of m.

### 1.2.1 Overview of Graphs

The growth of the social networks such as Orkut, Facebook, Instagram, LinkedIn, Twitter, Pinterest, Google+ etc. has rapidly increased in the last decade. These networks can be represented in the form of graphs. The most interesting and valuable objects in mathematics is graph. A graph or network is a collection of nodes or vertices where nodes or vertices can be associated together via edges or links. A vertex may exist in a graph but it is not the necessity that belongs to an edge.

The order of the graph is described as the number of vertices in a graph and the size of the graph is described as the number of edges in a graph. The degree of a vertex is described as the number of edges associated to it.

An edge is defined as pair of vertices ($a_1$, $a_2$) where $a_1$ and $a_2$ belongs to the vertex set. A graph may either be directed or undirected. A graph is said to be directed graph have edges with direction. Edge from $a_1$ to $a_2$ is not same as edge from $a_2$ to $a_1$. One-way relationship is indicated by the edges in which edge can be traversed in a single direction. In an undirected graph, edges do not have a direction. Edge from $a_1$ to $a_2$ is same as edge from $a_2$ to $a_1$. Two-way relationship is indicated by the edges in which each edge can be traversed in both directions.

A graph may either be weighted and unweighted. The weighted graph is defined as the weight or number that is associated with the edge. Weight assigned to the edges of a graph totally depends upon the type of the network for which it is used. A network defined its own weights. For instance a graph representing twitter data has tweets and retweets as the edge weights. On the other hand, there are unweighted graphs which have Boolean value associated with them where 0 represents that there is no edge and 1 represents the contrary.

### 1.2.2 Creating Social Media Network

As a result of several online transactions that happens everyday we can create social networks. A network can be created out of these transactions. Some entities are

related to each transaction like a tag or a photo. In a blogging network there can be entities like the blog article, blog commenter and the text. Therefore, a relation is made within the entities that are similar to each other which results in a network. Hypergraphs can be used for the representation of complex mathematical networks. It can handle multi way edges too. There are also some drawbacks of hypergraphs. One of them is that majority of the network analysis methods are not applicable to it. For this reason we now focus on simpler networks which are partial aspect of bigger and complex networks. Therefore, we prefer simplified networks that have one- or two-mode vertices which are simple and connected. Hence network is formed which is used for the social analysis.

Preprocessing is required before analysis on complex network. The process requires simplification and network cleaning. Fig 1.3 shows the pre-analysis steps of the network.

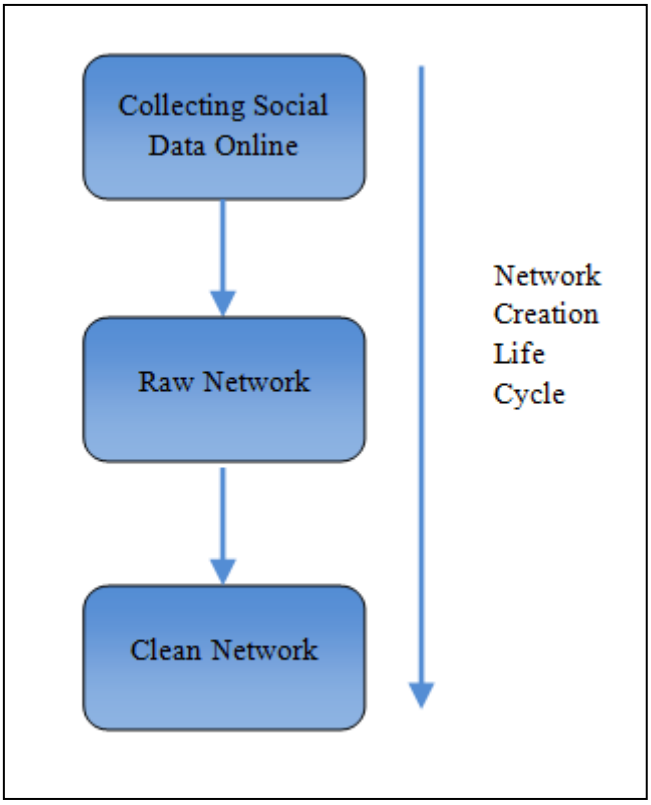


**Fig. 1.3: Pre-analysis Steps**

## 1.3 Community

There is a no unique definition of community as there are several related works and many adopted perspective. The definition of community is made related to the network structure of the entire framework which is under review and is related to some kind of property such as set of vertices, global definition of the networks etc. A community can also define the respective to its domain of study. Hence in this context we will first define the online network communities and then further we will connect this to establish system based definitions which are quantitative in nature.

At the highest level, taking a Social Media network G = (V, E), we can consider a social media community as a subgraph of the network consisting of a set $V_c \sqsubseteq V$ of Social Media entities which are related to each other with a common interest . The interest of study can range from topic real network, an event or circumstances for a cause. For instance, in a network where blogging is done all the blogger tag and comments that are associated to the domain of renewable energy consist of the respective community.

**Implicit Communities**

Implicit Communities are those communities which are already existed in the system and hold up to be discovered. The human exertion and consideration is not required for their creation is the important property which differ the implicit communities from any other communities.

**Explicit Communities**

Explicit Communities can be made on the basis of human consent and human decision. Flickr and Facebook are the examples of the Explicit Social Communities.

### 1.3.1 Community Structure and Attributes

Various definitions of community have been proposed but it is viewed as a set of vertices and relationship between the vertices from vertex set and defined though boolean decisions. The idea of community and community membership is considerably more convoluted in reality. In the previously mentioned community definition like Clique Percolation [7], the local ones [27, 8] and many others [9, 10],

there is a possibility for communities to overlap (Fig 1.4). Overlapping communities are those communities in which an attribute or an entity can occur in more than one community. Overlapping community is vital for social networks. For example same person can be part of two communities like family and friend.

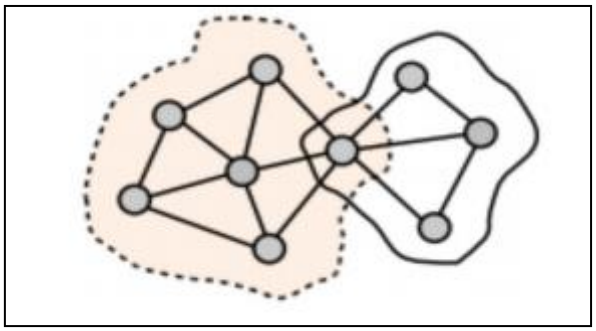

**Fig. 1.4: Overlapped Communities**

Also there are many attributes which can be related to vertices of a network. For example, there can be various vertices that may take part with numerous degrees in a community (Fig 1.5).

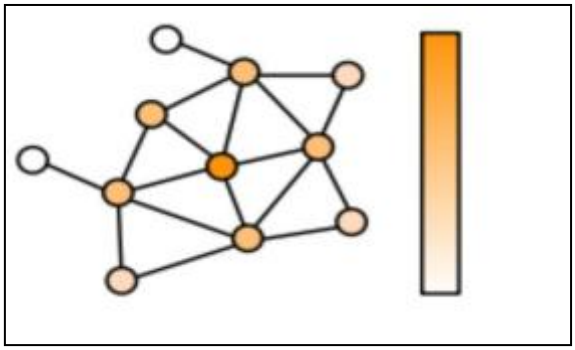

**Fig. 1.5: Weighted Members**

At last, there is a possibility of imposing multi-scale structure or hierarchical structure (Fig 1.6) on communities. The organization of community can be understood at different scale in a numerous system. For example, a group of users of web-based social networking can be collectively in a community which focused on a particular topic and simultaneously they can also be a part of broader community. In social media the multilevel of community which are organised do not include any type of hierarchical organization because of the impose constraint of the various levelled demonstrate that are excessively prohibitive for making the mode of emerging and uncontrolled nature of web-based social networking.

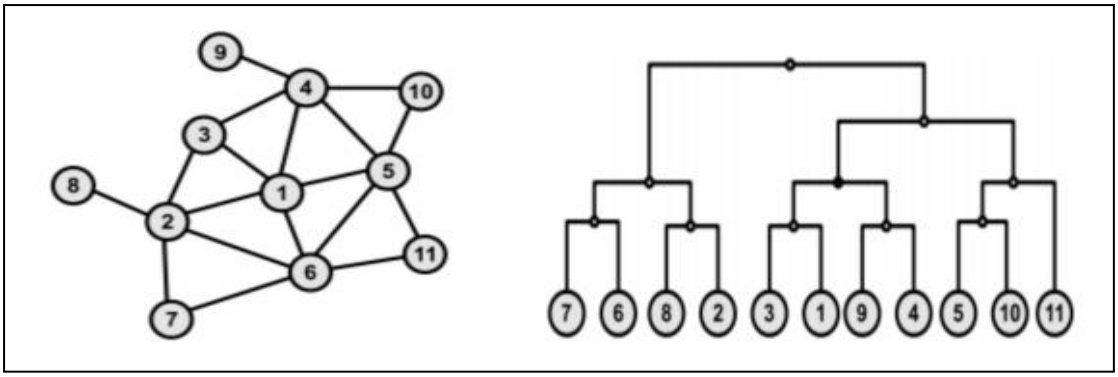


**Fig. 1.6: Hierarchical Structures**

## 1.4 Twitter: A Microblogging Service

Twitter is a social networking service which allows the user to post and read the short message of 140 characters known as "tweets". Twitter was developed in 2006. It allows the user to post their ideas, opinions or messages in defined number of words. 140 characters are sufficient to update the status via text messages. There are more than 1000 million users who have already registered on twitter account and produce 300 billion regularly. There are two types of users for twitter account. One is registered users who can only read the tweets and another are registered users who can read and post the tweets. It is a public platform for all the people of different age categories all over the world. Twitter has emerged for celebrities and business as a social media networking platform, not only for the average individual user. The opinion and behaviour of the individuals and groups can be understandable by analyzing the tweets [36]. Twitter provides the well documented and clean API for the consumption of data which is open publically.

Twitter allows one user to follow another user without mutual acceptance that why twitter follows the asymmetric mechanism. In the form of tweets the social digital data is being produced in a huge amount on regular basis. Data generated by twitter is heterogeneous in terms of content because user can post a text, image, video and audio in any format. The user's opinions can be analyzed in various fields such as politics, advertising, sports, marketing, education, business etc. by using this large social dataset. The opinions are differing from person to person in above mentioned fields; it can either be an appraisal or a criticism. Fig 1.7 shows some of tweets from different users about the "MCD Results in Delhi".

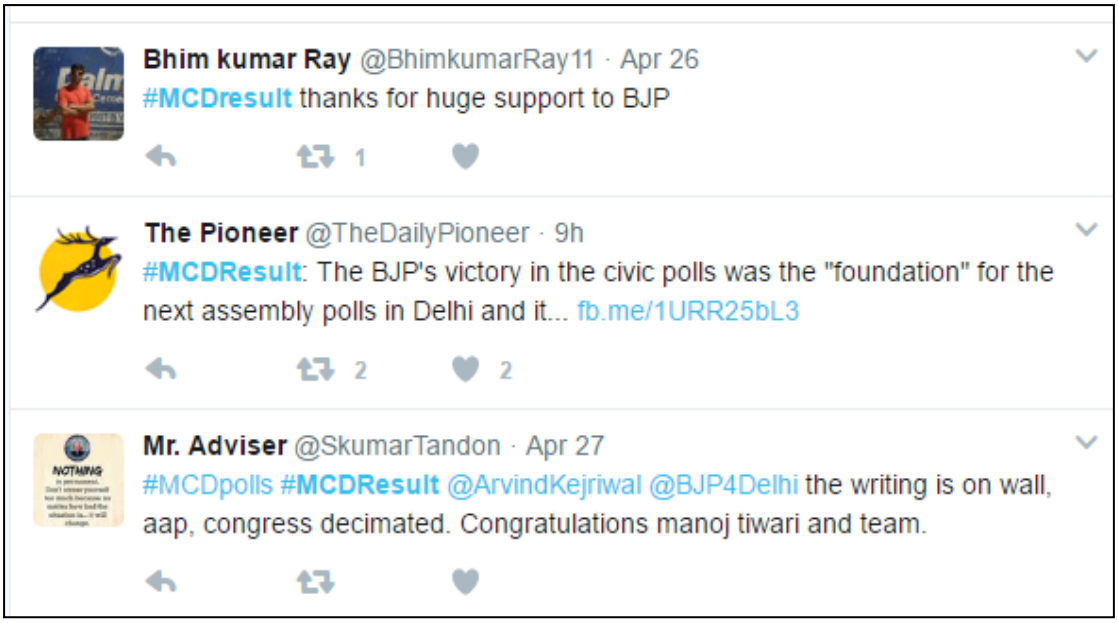


**Fig. 1.7: Tweets from Twitter**

## 1.5 LinkedIn: The Professionals Social Web

The business and professional relationship are focused by the powerful and popular social network site which is LinkedIn. LinkedIn is launched in 2003. It provides a platform to the individual so that they post their resume, search a job according to their education and professional qualification, connect with other individual of same professional qualification and recommend their friends. LinkedIn may appear like other social networking sites but the data provided by LinkedIn is different as the data provided by the other social networking sites like Facebook, Twitter etc. Currently, LinkedIn has 500 million members from 200 countries [38]. The site is available in 24 languages [39].

The data available at LinkedIn is very unique in relation to the data available at the other social networking sites because of the professional characteristics of LinkedIn. LinkedIn provides the large amount of data of the millions of the subscribers. This large amount of dataset is not able to handle or manage by the traditional database management techniques. The dataset become complex due to the number of subscribers has increase in multiple of thousands [13]. The dataset is very simple if it contains less than 2000 profiles. The researchers have proposed various tools and techniques to store compute and handle the large amount of dataset.

To manipulate these complex and huge datasets is a difficult task. This can be possible by dividing these datasets into small subsets by using the specific tools and methods for constructive outcomes. A lot of information can be collected about individuals and groups by investigating the LinkedIn data. The LinkedIn provides the information (school and college information, employer, profession, past history of job etc.) of the individuals to the users even they are not connected to the network [40]. The semi structured data is provided by the LinkedIn members. LinkedIn members freely give away job titles, industry information, location information, skill sets, educational qualification etc.

The behavior and properties of individuals and group can be understood by investigating the LinkedIn data. The professional relations of individuals and groups can be explored by analysis the LinkedIn data. The personality of the individual can be picturized by collecting all the LinkedIn relations, the previous background of the designation, occupation, locality, employer etc. Some parameters can be set for LinkedIn in order to compare individuals. Various types of professional communities can be detected by hierarchical clustering on the basis of job title and companies by analyzing the LinkedIn dataset.

# Chapter 2 LITERATURE SURVEY

## 2.1 Categories of Community Detection Methods

There are various community detection algorithms or methods that can be used to find the clusters in real world networks. The various algorithms to find the communities are discussed below:

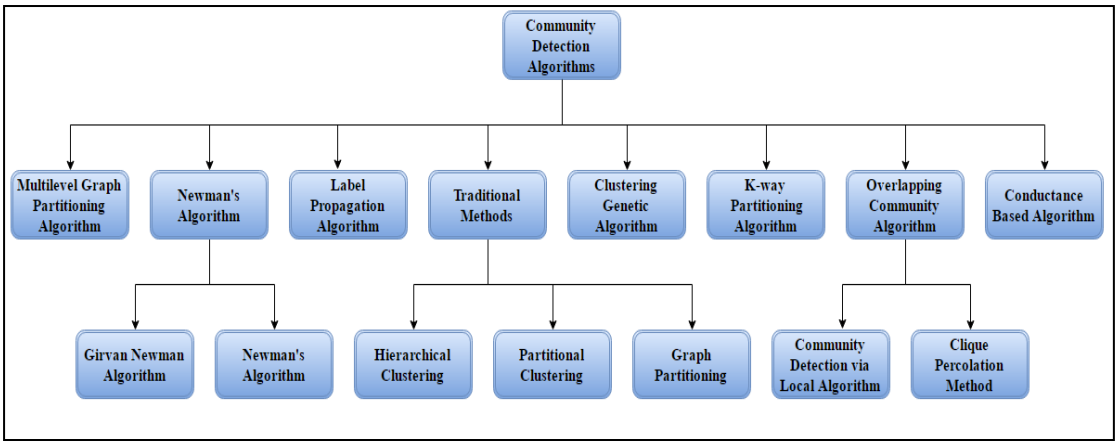


**Fig. 2.1: Algorithms for Community Detection**

### Graph Partitioning

Graph Partitioning is the problem of isolating the graph into groups of predefined size such that the number of edges between the groups is minimal. The cut size is the number of edges between the groups. Most of the graph partitioning method is based on the partitioning the graph into two separate groups iteratively: the spectral bisection method [11] which uses the properties of spectrum of Laplace Matrix and eigen vector of the graph and Kernighan-Lin [12] algorithm which optimize the community structure over partition of the graph in a greedy way.

The Kernighan-Lin (KL) algorithm is one of the oldest methods and is still frequently used to solve the graph partitioning problem. The algorithm tries to maximize the benefit function. The benefit function is the difference between the number of edges within the group and the number of edges between the groups. The algorithm works as: initially partition the graph into two groups of a predefined size by using some properties of the graph or randomly. The swapping of pair of vertices continues until the benefit function is reached maximum.

**Hierarchical Clustering**

Hierarchical clustering is deterministic and unsupervised clustering technique in which it processes the distance between the data items and stores them in a matrix [13]. Then traverse the whole matrix and merge or split the data items based on the similarity measure such as Jaccard distance, masi distance etc. The merge and split are determined in a greedy manner. The results of the hierarchical clustering are usually represented in a dendrogram. Hierarchical Clustering Algorithm is divided into two groups:

1. *Agglomerative Hierarchical Clustering Algorithm*
2. *Divisive Hierarchical Clustering Algorithm*

Both the algorithms are exactly opposite to each other. The agglomerative hierarchical clustering works as every data point is considered as its own cluster. The data points can be grouped together into a parent cluster based on the similarity measure between them. The process continues until all the data points are in one cluster. The cluster hierarchy generated by this algorithm is bottom up. That is why, it is also known as bottom up clustering. Fig 2.2 shows the schematic figure of Agglomerative Clustering.

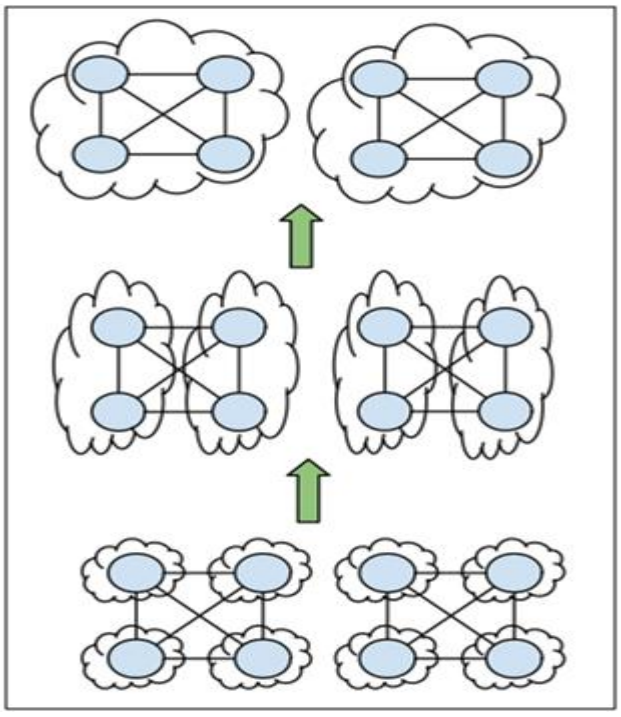

**Fig. 2.2: A schematic diagram of Agglomerative Clustering**

The top down cluster hierarchy is also generated by the divisive hierarchical clustering algorithm. The working of this algorithm is exactly opposite of the agglomerative hierarchical clustering algorithm. The algorithm works as that all data points are considered to be in one cluster. The data points from the cluster are split based on the similarity between them and make a sub clusters. The process continues until each data point is in its own cluster. The similarity between the clusters can be defined by the finding the minimum distance between the clusters, maximum distance between the clusters, average distance between the clusters, Jaccard distance and so on. Fig 2.3 shows the schematic figure of Divisive Clustering.

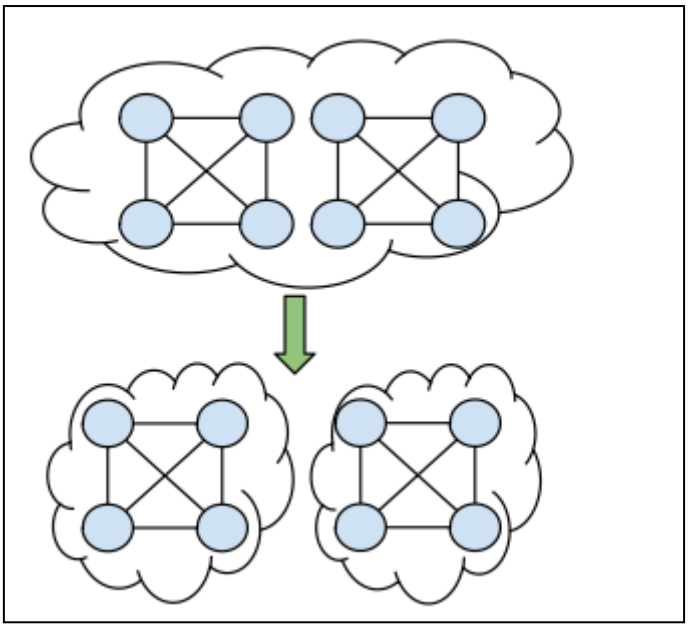

**Fig. 2.3: A schematic diagram of Divisive Clustering**

## Partitional Clustering

In Partitional clustering algorithm, a data can be separated into different clusters or groups by minimizing the objective function. It is also known as centroid based clustering [14]. A document is compared with its nearest neighbour or the mean of the documents in order to minimize the function of the associated document with that cluster.

K-means [15] clustering algorithm is an unsupervised learning algorithm which is used for the unlabelled data i.e. data are not labelled into any group or cluster. The objective of this algorithm is to find the clusters in the data with the already given number of clusters. The number of clusters and the dataset are the inputs of the

algorithm. The dataset is the collection of data for each data point. The number of clusters can either be randomly selected or randomly generated from the dataset. The algorithm works as: firstly initialize the number of clusters and the set the centroid of the clusters. Each data point is assigned to a cluster based on the smallest distance between the centroid and the data point. The centroids are updated or recomputed by taking the average of the data point assigned to the cluster. The process continues until the stopping criterion is met. The stopping criterion is any one of them which are data points is not changing the clusters, the sum of distances is minimized or the number of iterations are reached maximum.

**Girvan Newman Algorithm**

Girvan Newman Algorithm [16] [17] is the most well-known Divisive Algorithm. The idea behind this algorithm is to find the good inter community edges and those edges are the candidate for being "between" communities. The fundamental concept is called edge betweenness which is described as the number of shortest part between the pair of nodes or vertices that keep running along the edge. An equal weight is given to each path, if there are various paths between a pair of nodes or vertices. The edges with the highest edge betweenness value are the edges between the communities. The algorithm removes the edges until no edge is leftover. The edge with the highest edge betweenness is removed at each iteration.

**Newman's Algorithm**

Newman proposed the Newman's Algorithm [18] in 2004 and falls in the category of agglomerative hierarchical clustering [19] [20]. The modularity function which is the fraction of edges that is present inside the communities or group minus the expected value of the same quantity if the edges are all over the graph.

The algorithm begins with each and every vertex has its own community or group. At each iteration, the algorithm joins the pair of vertices or communities that results the smallest decrease or greatest increase in the modularity. The order of the join can be shown by the "dendrogram". The different partitions into communities can be given by cut through the dendrogram at various levels. The maximal value of modularity is selected for the best cut. The process of joining the pair of communities continues until there is no edge left which increase the modularity.

**Label Propagation Algorithm**

Raghavan et al [21] proposed the Label Propagation Algorithm. For online social network, it is a good candidate for used as a community detection algorithm because it works in near real time. It requires neither the prior information about the communities nor predefined objective function. It is computationally less expensive than the other algorithms because of its low time complexity. The algorithm uses the greedy approach and does not provide the optimal solution. For every iteration, a result may change.

The idea behind this algorithm is to initialize all nodes with a unique label and then a particular node searches the maximal occurring label in its neighbourhood and changes to that label. If the number of labels occurring in the neighbourhood are same for the different label of a node; then, tie is broken randomly. This process continues until no node is left to change the label. The nodes with the same label make a community.

**Multilevel Graph Partitioning Algorithm**

Multilevel Graph Partitioning Algorithm is proposed by Karypis et al [22] is based on the graph partitioning algorithm in which the size of graph is reduced by shattering the vertices and edges, partition the smaller graph and after that uncoarsen it to construct a partition for the initial graph. This algorithm consists of three phases for bisecting a graph:

*Coarsening Phase:* During the coarsening phase, the initial graph is partitioned into associated subgraphs and contracting them to single vertices.

*Partitioning Phase:* The second phase of multilevel graph partitioning algorithm is to partition the graph into two sections using any partitioning method (for example, Kernighan-Lin algorithm, Spectral Partitioning).

*Uncoarsening Phase:* During the uncoarsening phase, the partition of the graph is anticipated to the original graph which includes replacing each vertex by the subgraph that was shrinking to that vertex during the coarsening phase.

**Conductance Based Community Detection Algorithm**

Lu et al. [23] proposed the Conductance Based Community Detection Algorithm for the weighted networks. The algorithm works as: first find the two nodes whose edge weight is highest and considered those nodes in one community and conductance is calculated. In the expanding process, nodes which are adjacent to community are found and calculate the belonging degree of those nodes. After that choose the node whose belonging degree is highest and merge with the community to form a new community. Then again calculate the conductance of the new community and compare the conductance of the new community with the conductance of the older community. If the conductance of the new community is less than the conductance of the older community then the expanding process will be continued for the new community as well; otherwise the older community is designated as a the community. The edge which is in the community is removed from the edges set and the process is repeated till no edge is left in the set.

**K-Way Partitioning Algorithm**

The community structure of the network is better, if the value of the modularity is higher. But to optimize the value of modularity, a good amount of time consumed in community detection process. Thus, k-way partitioning algorithm is proposed by Behara et al. [24] which has less time complexity as compared to other community detection algorithms.

This algorithm works as: the network is partitioned into the specified number of communities by assigning each vertex or node of the network into one of the communities. After that, calculate the modularity matrix of the given network and also calculate the eigen vector corresponding to maximal eigen value of modularity matrix. Then randomly assign the values in vector to each node of the network and put the node into one of its communities. The value of eigen vector which is assigned to each node of the network is to be normalized by the number of communities.

### Clique Percolation Method

Palla et al. [25] proposed the Clique Percolation Method in 2005. This method is used to identify the overlapping communities in a network. The cliques are formed by the internal edges of a community because of the high density. The basic idea of this method is firstly detects communities of size k in a network and creates a clique graph. Then add edges if two cliques share k-1 common nodes and communities are represented by the connected component.

### Overlapping communities detection algorithm via local algorithm

There is another method to detect the overlapping community in a network based on the node strength and is proposed by Chen et al. [26]. There are two main components of this method: finding the initial community and expanding the community.

In the first component, first find the node strength of each node and then select the node with the highest node strength and find its neighbors. These nodes make an initial community. After that, calculate the belonging degree of each node which is present in the community. If the belonging degree of each node in the community is less than the threshold belonging degree then remove that node from the community. Repeat this step until the belonging degree of each node is greater than the threshold belonging degree and obtain the initial partial community.

In the second component, find the neigbhors of the initial community and calculate the belonging degree of each neigbhors. Add all the neighboring nodes directly into the community, if the belonging degree of the node is larger than the pre-specified belonging degree. Repeat both the components to find all other communities.

Table 2.1 shows the comparison of various community detection methods.

**Table 2.1: Review of Community Detection Algorithms**

| Algorithm | | Pros | Cons | Fitness / Objective Function |
|---|---|---|---|---|
| Clustering | K-Means | 1. Easy to implement<br>2. Reasonable performance | 1. Need to specify the size of the clusters beforehand<br>2. Often ceases at a local optimum | Mean Distance |
| | Hierarchical Clustering | No need to specify the size of the clusters beforehand | 1. Do not know where to cut the dendrogram tree<br>2. May get bad results if the merging heuristic is not good | Euclidean Distance |
| Newman's Algorithm | Girvan Newman algorithm | No need to specify the size of the clusters beforehand | Slow | Edge-Betweenness |
| | Newman's Algorithm | 1. Faster than Girvan Newman algorithm<br>2. No need to define the size of the clusters ahead of time<br>3. May get good partitions | No theoretical guarantee contrast with the greedy algorithm | Modularity |
| Graph Partition | Kernighan-Lin | Fast | Need to specify the size of the clusters beforehand | Benefit Profit |
| Label Propagation Algorithm | | 1. Time complexity is less<br>2. Efficient<br>3. Can detect overlapping community | Can detect one community | ----- |

| Algorithm | | Pros | Cons | Fitness / Objective Function |
|---|---|---|---|---|
| Conductance Based Community Detection Algorithm | | 1.Can detect overlapping community<br>2. Efficient | ----- | Conductance and Belonging Degree |
| K-way Partitioning Algorithm | | Low time complexity | Need to specify the size of the clusters beforehand | Modularity |
| Multilevel Graph Partitioning Algorithm | | ----- | Unbalanced Partitions | Modularity |
| Overlapping Community Detection | Clique Percolation Method | Can detect overlapping community | 1. Good for networks with many full connected subgraph<br>2. Fail to give relevant covers for graph with few cliques | ----- |
| | Community Detection via local algorithm | 1. Able to detect overlapping community<br>2. Work with Weighted Graph | ----- | Belonging Degree, Node Strength, Edge Betweenness |

## 2.2 Measures of Community Quality

The above section discusses the different community detection algorithms. But to differentiate the different communities either good or bad, there are various measures and are illustrated below:

1. ***Modularity:*** Modularity is defined as to measure the quality of partition of a network into groups. Network has high modularity if there is a strong connection between the nodes within community and weak connection between the nodes with different community. The modularity is defined by Clauset et al. [27] is:

$$Q = \frac{1}{2n}\sum_{pq}\left(B_{pq} - \frac{k_p k_q}{2n}\right)\delta(p,q) \quad (1)$$

where $B_{pq}$ is the adjacency matrix, n is the number of edges, $k_p$, $k_q$ are the strength or degree of nodes, $\delta$(p,q) is a function which returns 1 if p and q are in same community otherwise 0.

2. ***Edge Betweenness****:* Edge Betweenness of an edge is defined as the number of shortest path between pair of nodes that keep running along it [16]. Edge Betweenness is defined by Girvan and Newman is :

$$\mathrm{EB(e)} = \sum_{\mathrm{v_i}}\sum_{\mathrm{v_j}} \frac{\sigma_{\mathrm{v_i v_j}}(\mathrm{e})}{\sigma_{\mathrm{v_i v_j}}} \quad (2)$$

where $\sum_{v_i}\sum_{v_j}\sigma_{v_i v_j}(e)$ is the number of shortest path between $v_i$ and $v_j$ go through edge e and $\sigma_{\mathrm{v_i v_j}}$ is the total number of shortest path.

3. ***Conductance:*** Conductance is defined as ratio of total number or weight of edges connected with the community to the total number or weight of edges connected with the community and connected with other community. Conductance in the network is defined as [28] :

$$\emptyset(\mathrm{C}) = \frac{\mathrm{cut}(\mathrm{C_i}, \mathrm{C_j})}{\mathrm{W_c}} \quad (3)$$

where cut($C_i$,$C_j$) is weight of the cut edge of communities and Wc is the weight of all the edges in the community including cut edge.

4. ***Belonging Degree:*** Belonging Degree [26] is defined as ratio of aggregate weight or number of edges connected with the node with a community to the aggregate weight or number of edges associated with the node.

$$B(s, C) = \frac{\sum_{s \in C} w_{st}}{k_s} \quad (4)$$

where $w_{st}$ is the weight of an edge between node s and node t and node s belongs to the community C, $k_s$ is the node strength of an node.

5. ***Node Strength:*** Node strength [26] is the aggregate of weights of edges associated with the node. The node strength $k_u$ of a node is defined as

$$k_u = \sum_{v \in V} w_{uv} \quad (5)$$

where $w_{uv}$ is the weight of the edge.

6. ***Clustering Coefficient:*** Clustering Coefficient is defined as the probability that the neighbouring nodes of a node are associated. The following equation shows the clustering coefficient of a node in a network [29]:

$$C = \frac{2|e_{st}|}{k_q(k_q - 1)} \quad (6)$$

where $e_{st}$ is an edge between s and t node, $k_q$ is node strength of node.

# Chapter 3 RESEARCH PROBLEM

## 3.1 Problem Statement

It is interesting to note how the communication networks around the world have evolved around the world in the last decade and in later years. An important constituent of these networks is the Social Media, which is increasingly becoming the first choice among internet users these days when it comes to interaction with friends, relatives, colleagues etc. This is driven by the fact that we all have an inherent tendency to have know-how about the rest of the world and also sharing news and updates about ourselves and our near and dear ones. The importance of Social Media can be gauged by the fact that approximately two billion people over the world are its active users. They not only use it to exchange information and day to day updates about themselves and others and what's happening around, but also to accomplish various other goals like advertising, citizen reporting, content reviewing etc. Social Media thus provides a deep insight into the activities and behavioral patterns of user bases and has an immense potential of tapping this information to provide useful input for decision making in businesses and even government organs.

What's useful here is to realize this fact and devise strategies around the data so obtained after filtering and mining and converting such raw data into meaningful figures for decision making. A lot is still to be done in this area, but what's comforting that the work on it has already begun. Since the data is huge, and most of the times, repetitive, therefore it is necessary first of all to invent tools and techniques to make some sense of such data. Once the raw data has been successfully converted to meaningful sets of data, such data can be thoroughly analyzed to identify and prepare the pieces of information which are useful as per business needs and then deploying the same towards decision making in organizations.

Users have plethora of options to choose from in terms of what they can do on Social Media sites today. In addition to creating a profile and establishing communication with friends and family by posting updates and reading about others, they can create and join groups, create pages, do advertising, get user reviews, join a cause, and even vote on certain issues. This lays the foundation for a complex structure of

communication networks which in turn is a rich source of information for identifying user behavior and choices on internet. This process can be made easier by devising such software tools that serve as a generic option for any kind of complex and dynamic data structure obtained from such sites. Once this is done, it can now be used to making important business decisions and also predicting about future events. The social media researchers of today are engaged in observing the structure and association of various social networking websites.

One of the main research domain in social web analytics in studying the privacy, security, compactness and centrality of the user. Various trends can be identifying by analyzing the social web and various perspective of the society in different domains. Machine learning and other statistical tools can be used for the social data analysis. Using these tools, the challenges of getting effective results can be overcome.

## 3.2 Research Gaps

- A lot of research or study is going out in the field of social web analytics. Social networks are made by the analysis of daily basis communication among the users. These studies contain individual relationship related to other parameters.
- Normalization of data in preprocessing step is carried out to make the social data suitable for the analysis.
- A lot of research has been proposed on data clustering but there is a lack of research to analyze and utilize the human behaviour by using various clustering techniques on the social data such as LinkedIn, Facebook, and Twitter etc.
- Every social dataset contain the information of user's location and location where the user post the status or send messages. A very few research work is carried out in the field of social data analytics which used the spatial data present in the social dataset.
- Geoclustering is the one of the most important analysis and visualization of spatial data. An another main concern in the field of social web analytics is frequency analysis and visualization of the analyzed data like how many posts,

retweets and likes for their particular posts, tweets and status updates respectively.

- Clustering techniques for varied online communities according to parameters like location, company and job.
- Also we will study frequency analysis using the preprocessed and extracted dataset which contain tweets.

## 3.3 Research Objectives

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

- To study various data mining tools and techniques available for analysis and visualization of social data.
- Performance comparison of various existing Community Detection algorithms on different dataset using R tool.
- Removal of various redundancies present in the extracted social dataset using normalization or preprocessing techniques, to make the data suitable for the analysis.
- To perform frequency analysis on Twitter data and visualize the results in a graphical form.
- To develop a geo-clustering application for Twitter data using k-means clustering technique with R tool.
- To analyze and visualize the Professional LinkedIn data using Hierarchical clustering technique using Python.

## 3.4 Research Methodology

In our research work we will use python and R language. R language is used for the data acquisition, preprocessing and visualization of the clustered data. Python is an excellent scripting language which is used for manipulation of text. The various packages are available in R to analyze and visualize the social data. We can download the different packages in R using CRAN mirror. R framework is needed with R studio. The python packages can be downloaded and installed using *pip*.

**Community Detection**

- To get hands on experience with R and R tool.
- Download the data from different websites.
- Apply the different community detection algorithms on different dataset.
- Compare the applied algorithms based on the modularity.

T**witter**

- To get hands on experience with R and R tool.
- To get a Twitter developer account
- To authorize connection with twitter.
- To extract and preprocess the tweets from twitter for further analysis.
- To perform the k means cluster analysis on the extracted data
- To perform the frequency analysis of tweets
- To visualize the analyzed data

**LinkedIn**

- To get hands on experience with python.
- To export the data connections in CSV format.
- To normalize the extracted data for further analysis.
- To perform the frequency analysis of company names and job titles.
- To perform the Hierarchical cluster analysis on the extracted data.
- To visualize the analyzed data

# CHAPTER 4 A COMPARATIVE ANALYSIS OF COMMUNITY DETECTION ALGORITHMS

## 4.1 Community Detection Methods in R

There are various packages for community detection in R. The most important package in R for community detection is 'igraph'. The package contains various community detection algorithms which are given below:

1. **edge.betweenness.community:** The idea of this method is to find the edge betweenness of the considerable number of edges in the network and aftwerward expel the edge with the most astounding edge betweenness. Further, again calculate the edge betweenneess of the edges which are affected by evacuating the edge with the highest edge betweenness. This process continues until no edge is leftover. This method is based on the Girvan Newman's Algorithm [16].

2. **fast.greedy.community:** The method is based on the modularity optimization. The idea behind this algorithm is that each node is treated as an individual cluster or community. The clusters or communities having most prominent increment or littlest lessening in modularity can be merged. This process continues until no cluster or community pair leads towards ost noteworthy increment or littlest lessening in modularity. The algorithm of this method is proposed by Clauset et al. [27].

3. **walktrap.community:** The method is to find the community on the basis of different distance measure between vertices depending on random walk [30]. The idea behind this algorithm is that each node or vertex is treated as an individual cluster. Then, the shortest distance between the clusters is found on the basis of some probability (likelihood of going from one vertex to another in some steps). The algorithm of this method was proposed by Pon et al. [30].

4. **spinglass.community:** This method is based on the Potts model [31]. Each node can be in one of the communities; the edges of the network specify the

nodes that would remain in the same or different communities. The process continues for a given number of steps, and at the end the communities are detected. Reichardt et al. [23] proposed the algorithm of this method.

5. **label.propagation.community:** The idea behind this method is to initialize all nodes with a unique label and then a particular node searches the maximal occurring label in its neighbourhood and changes to that label. If the number of labels occurring in the neighbourhood are same for the different label of a node; then, tie is broken randomly. This process continues until no node is left to change the label. The nodes with a similar label make a community. Raghavan et al. [21] proposed the algorithm of this method.

6. **leading.eigenvector.community:** It is based on the hierarchical top down approach that optimizes the modularity function by using eigen vector of the modularity matrix. The approach of the algorithm is to find the eigenvector of the modularity matrix. The network (graph) is partitioned into two parts such that modularity improvement is maximised based on the leading eigenvector. Further, the modularity contribution is calculated at each step in the subdivision of a network. The process stops when the value of the modularity can't be expanded more. Newman proposed the algorithm of this method [32].

7. **cluster_louvain:** It is a hierarchical and greedy approach based on the measure of modularity. Initially, each node is treated as an individual community. Calculation of modularity of each node is node is done. The node is further moved to the community with which it accomplishes the highest contribution to modularity. At the point, when no nodes can be reassigned, each node is considered an individual community, and the process continues with the merged communities. The process stops when the modularity cannot be expanded any more in a particular step.

## 4.2 Datasets

The performance of various algorithms for detecting communities is compared in terms of modularity by using four different data sets of real world networks. The dataset has been accessed from UM Personal World Wide Web Server and Stanford Large Network Dataset Collection. The real-world networks which are used for comparison of various approaches are Zachary's Karate Club, Bottlenose Dolphins, American College Football network and Facebook. Zachary's Karate Club is mostly used as a community detection networks.

**Zachary Karate Club Network:** Zachary Karate Club network [33] is a well known graph used to test the various community detection algorithms. The member of the Karate Club in the United States is shown by the nodes which are 34 in number and the individuals are connected with the edges.

**Dolphins Social Network:** Lusseeau et al. [34] describe the social network of a community of bottlenose dolphins living in Doubtful Sound, New Zealand. The network comprised of 62 dolphins represented by the nodes or vertices and edges represent those dolphins that were seen together more frequently.

**American College Football:** The next network is used for the comparison of various community detection algorithms is American College Football [16]. For the 2000 season, the schedule of Division I games is represented by the network. The team is represented by the nodes or the vertices of the network and the edges represents the usual season games between two teams.

**Facebook:** Facebook is the most popular social networking website that consists of millions of users. The data set comprised of 4039 nodes and 88234 edges. The nodes represent the individuals in the Facebook and the edges represent the connection between the individuals.

The details of the datasets used for the comparison of various community detection algorithms are listed in Table 4.1.

**Table 4.1: Datasets Details**

| Dataset | Number of Nodes | Number of Edges |
|---|---|---|
| Zachary Karate Club | 34 | 78 |
| Dolphin Social Network | 62 | 158 |
| American College Football | 115 | 613 |
| Facebook | 4039 | 88234 |

## 4.3 Results and Discussion

The Table 4.2 shows the modularity of different community detection algorithms for different datasets. As we know, the value of the modularity gives the tightness between the nodes in the communities. As we shown in the table, the modularity gained by the Walk Trap Algorithm is the smallest followed by the Label Propagation Algorithm and the highest modularity is gained by the Spin Glass for the Zachary Karate Club dataset. Similarly, for Dolphin dataset the smallest modularity is gained by the Label Propagation and highest modularity is again gained by the Spin Glass algorithm. For American College Football dataset, the smallest modularity is gained by the Leading Eigen and highest by the Louvian Algorithm. The smallest modularity is gained by the Fast Greedy algorithm and highest gained by the Louvian Algorithm. The value of modularity is not defined by the edge betweenness for the Facebook dataset because this algorithm is not working for the large dataset [35].

**Table 4.2: Modularity of various algorithms for different datasets**

| Algorithms | Karate | Dolphin | Football | Facebook |
|---|---|---|---|---|
| Edge Betweenness | 0.401298 | 0.519382 | 0.599629 | Not Working [35] |
| Label Propagation | 0.374424 | 0.373482 | 0.568772 | 0.801409 |
| Walk Trap | 0.353221 | 0.488845 | 0.602914 | 0.81194 |
| Fast Greedy | 0.380670 | 0.495490 | 0.549740 | 0.777380 |
| Spin Glass | 0.419789 | 0.528519 | 0.598158 | 0.833468 |
| Leading Eigen | 0.393408 | 0.491198 | 0.492605 | 0.799134 |
| Louvian | 0.418803 | 0.518531 | 0.604569 | 0.834786 |

The graphical representation of the modularity of different approaches for different dataset is shown below. With this representation, we identified which algorithm works well for which dataset. Fig 4.1 shows the modularity obtained using different

algorithms for different datasets. X-axis represents different datasets and Y-axis represents modularity. From Fig II., it can be observed that the Spin Glass and Louvain Algorithms have highest modularity for karate club, Spin Glass has highest modularity for Dolphin dataset, Walk Trap and Louvain have the highest modularity for Football dataset, and Louvain have the highest modularity for Facebook dataset.

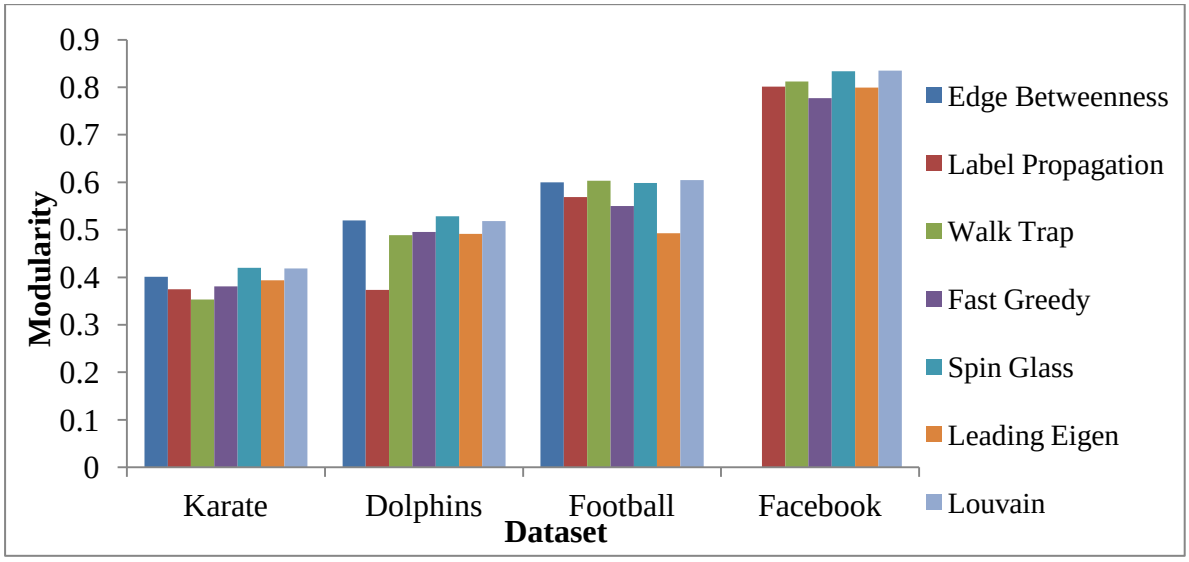


**Fig. 4.1: Comparison of various algorithms for different datasets based on modularity**

Table 4.3 represents the clustering coefficient of different datasets. It represents the overall level of the clustering in a network. As shown in a table, the clustering coefficient is small for the Zachary Karate Club followed by the Dolphins dataset and Football dataset. The highest clustering coefficient is for the Facebook dataset which means that the nodes in a network are strongly connected as compared to the other datasets and has high tendency to make a cluster.

**Table 4.3: Clustering Coefficient of different datasets**

| **Dataset** | **Clustering Coefficient** |
|---|---|
| Karate Club | 0.2556818 |
| Dolphins | 0.3087757 |
| Football | 0.4072398 |
| Facebook | 0.5191743 |

The graphical representation of the clustering coefficient of different dataset is shown Fig 4.2. With this representation, we easily identified which dataset has the highest tendency to make a cluster. X axis represents the different dataset taken for the

comparison of various community detection algorithms and the Y axis represents the clustering coefficient of the different dataset.

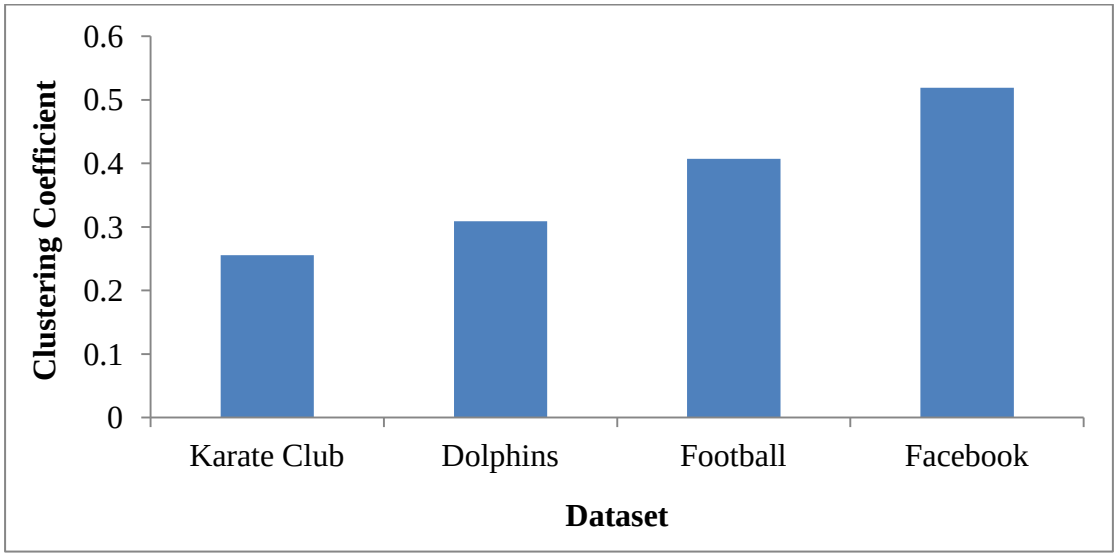


**Fig. 4.2: Clustering Coefficient of various datasets**

# CHAPTER 5 ANALYSIS AND VISUALIZATION OF TWITTER DATA

Twitter is a social networking service which allows the user to post and read the short message of 140 characters known as "tweets". Twitter was developed in 2006. It allows the user to post their ideas, opinions or messages in defined number of words. 140 characters are sufficient to update the status via text messages. There are more than 1000 million users who have already registered on twitter account and produce 300 billion regularly. There are two types of users for twitter account. One is registered users who can only read the tweets and another are registered users who can read and post the tweets. It is a public platform for all the people of different age categories all over the world. Twitter has emerged for celebrities and business as a social media networking platform, not only for the average individual user. The opinion and behaviour of the individuals and groups can be understandable by analyzing the tweets [36]. Twitter provides the well documented and clean API for the consumption of data which is open publically.

Twitter allows one user to follow another user without mutual acceptance that why twitter follows the asymmetric mechanism. In the form of tweets the social digital data is being produced in a huge amount on regular basis. Data generated by twitter is heterogeneous in terms of content because user can post a text, image, video and audio in any format. The user's opinions can be analyzed in various fields such as politics, advertising, sports, marketing, education, business etc. by using this large social dataset. The opinions are differing from person to person in above mentioned fields; it can either be an appraisal or a criticism. Fig 5.1 shows some of tweets from different users about the "MCD Results in Delhi".

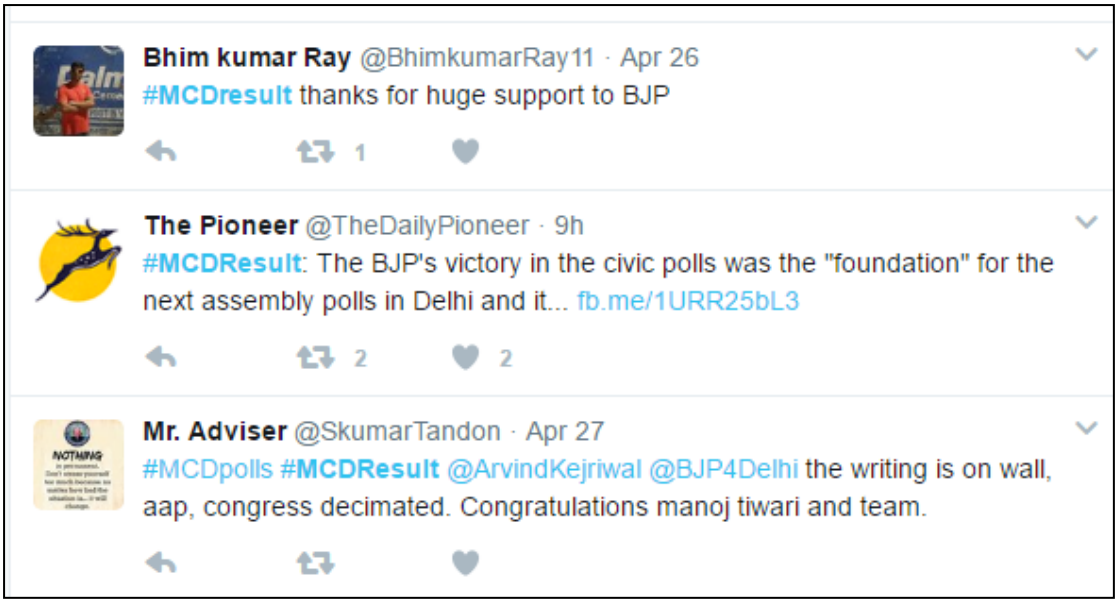


**Fig 5.1: Tweets from Twitter**

## 5.1 Implementation Methodology

API stands for Application Programming Interface which defines the set of methods to communicate between various software components. The twitter data such as followers, tweets, retweets, latitude, longitude etc are extracted by using different twitter APIs. The *REST APIs* helps the developer or programmer to read, write and access the twitter data. The *Streaming APIs* provides the uninterrupted access to the twitter data for a search query. It runs continuously until the internet connection interrupts or they kill. Twitter provides the different streaming endpoints for accessing the twitter data.

*Public Streams:* Twitter data which is publically available can be accessed by using these streams. These streams are suitable for particular user, particular research interest or mining the data.

*User Streams:* The data which is related to specific user can be accessed by using these streams. By using these streams we can access followers, tweets or retweets.

*Site Streams:* These streams are used for those servers which are linked to twitter on favor of many users.

To access server resources on the behalf OAuth 1.0a authorization mechanism is used by Twitter to provide its data for development purpose. Fig 5.2 shows the implementation methodology.

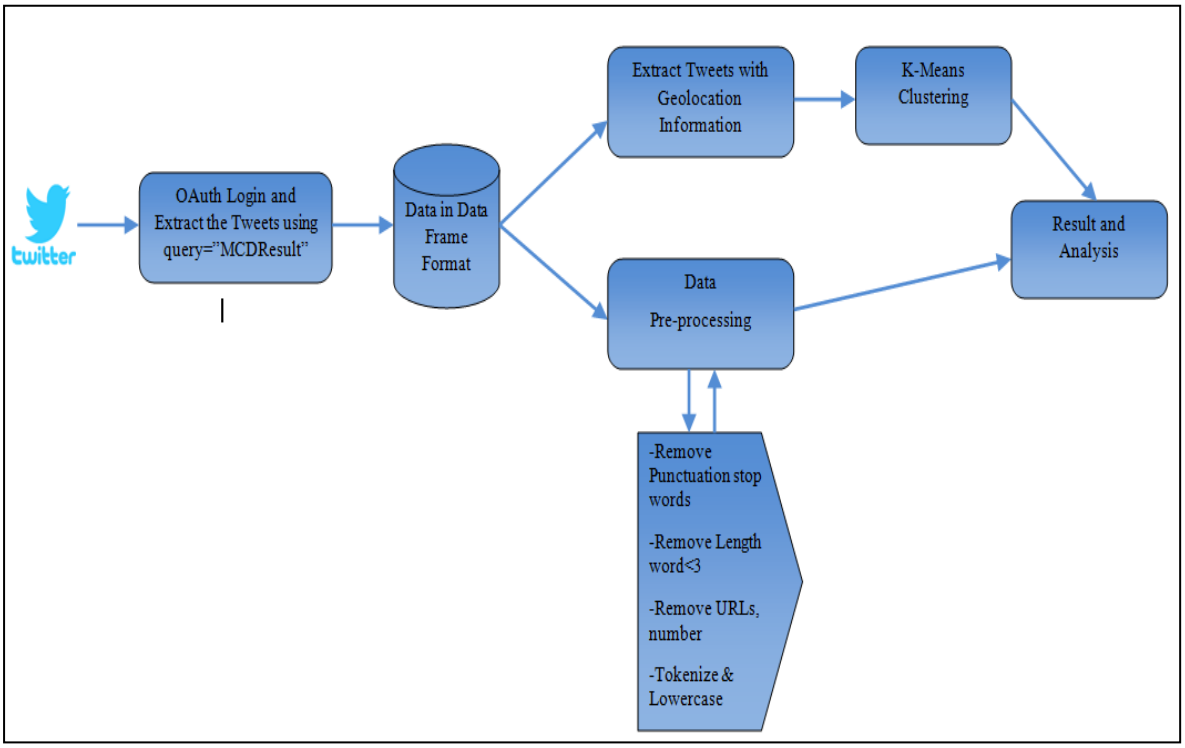


**Fig. 5.2: Methodology of Proposed Model**

In first step, authorization credentials obtained by creating and registering for an application on Twitter development environment and R twitter packages, are used to access the Twitter REST API. MCD results tweets are extracted with the parameter "#MCDResults" by using the search query. All extracted tweets are in the Data Frame. The extracted tweets contain the noisy data and the data needs to be clean for the analysis and also tweets which are unique and contains the geographic coordinates are considered. Further in next step, tweets are visualized in the word cloud. Also, the K-Means Clustering algorithm is applied to cluster the twitter data according to the geographic coordinates. The geographic clustered is visualized on Google Map using R language.

The various packages are available in R to analyze and visualize the twitter data. We can download the different packages in R using CRAN mirror. Table 5.1 describe the purpose of various packages used for analysis and visualization of twitter data.

**Table 5.1: Various Packages in R for Analysis and Visualization**

| R Packages | Description |
|---|---|
| twitterR | Interface is provided to the Twitter web API |
| data.table | Used for working with large dataset |
| ROAuth | Interface is provided by R OAuth to OAuth 1.0 |
| Leaflet | Interactive map is created and customised |
| Rtweets | Collecting the Twitter data via communicating with Twitter APIs |
| Maps | Display of maps |
| ggplot2 | Improve the quality of graph |
| Gdata | Used for data manipulation |
| Ggmap | Used to access the maps from the Google Map API |
| Rio | Import and export the data in different file format |
| Plotly | Translate the graph into interactive web based version |
| RColorBrewer | Provide the color scheme for graphs and maps |
| MASS | Functions and datasets is provided to support Venables and Ripley |
| RCurl | Allow to upload or download the file from web servers, password authentication, handle redirect etc. |
| Tm | Provide a framework for text mining |
| wordcloud | Used for making a cloud of words based on the count of words |
| SnowballC | Collapsing words to a common root to aid comparison of vocabulary |

## 5.2 Twitter Mining Application Setup and Data Extraction

We need to create Twitter Application to initiate authorized calls to REST API and collect the data for analysis and visualization phase. Without sharing the credentials to access the data from various services OAuth can be used as an authentication protocol [37].

### 5.2.1 Creating Twitter application

Firstly, we will need to create the new application at *https://dev.twitter.com/apps* from Twitter Application Management panel by filling the required fields for application to get the access token.


**Fig. 5.3: Twitter Application Setup**

### 5.2.2 Obtaining Twitter Credentials

Under the Keys and Access Tokens tab, there are four credentials are located which are Consumer Key (API Key) and Consumer Secret (API Secret), Access Token and Access Token Secret for the development and API access. These credentials provide to authorize the application and make the API requests on the behalf of its Owner. Fig 5.4 shows Consumer Key and Consumer Secret on the Application Settings.

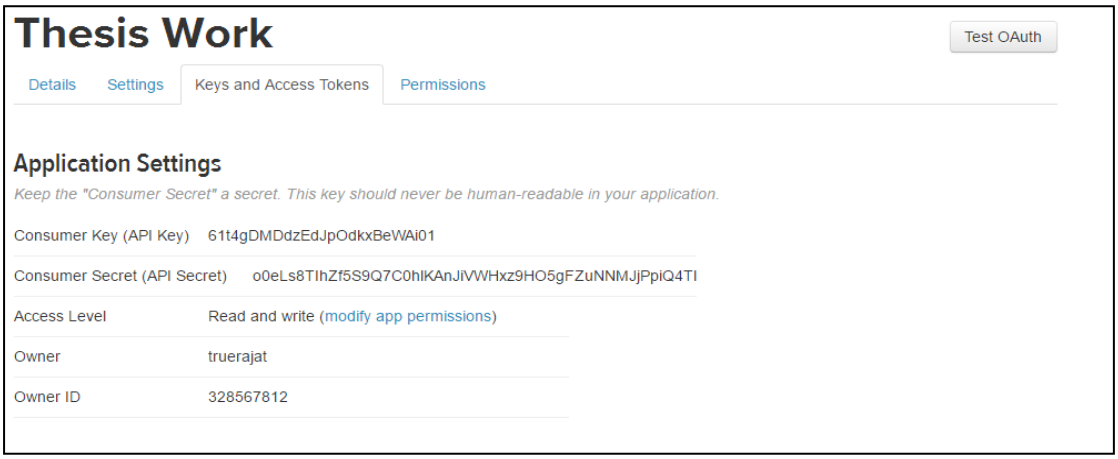


**Fig. 5.4: Consumer Keys**

Access tokens are used to make API request on the behalf of owner. The access level permission settings of application can be changed if necessary because it set to Read and write for this purpose. We also regenerate or revoke the tokens as shown in Fig 5.5.

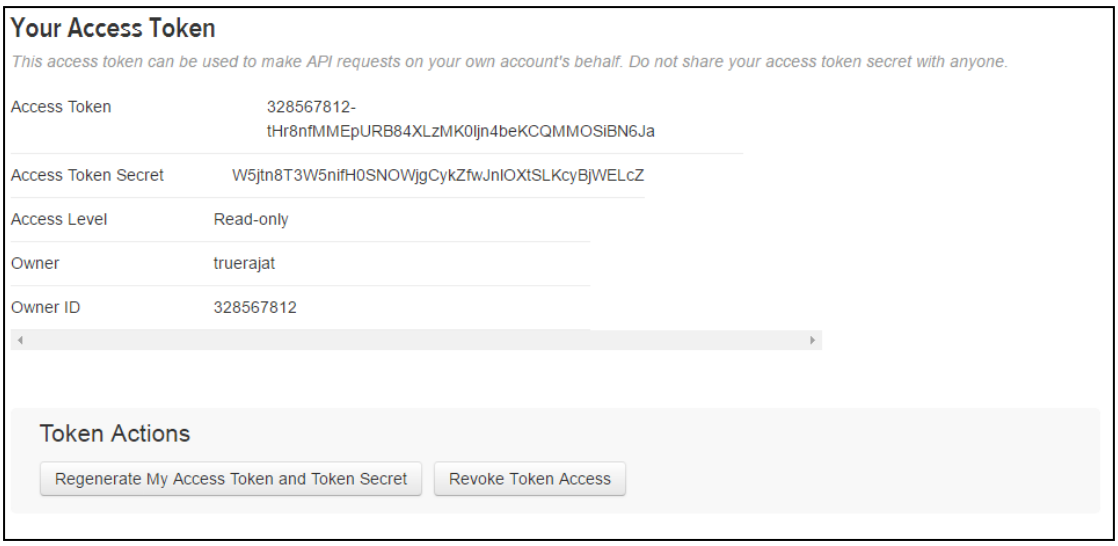


**Fig. 5.5: Access Token Credential Settings**

### 5.2.3 Creating Connection

After creating an application and obtained credentials successfully for API, initialize the connection and collect the sample tweets in order to analyse each tweet. To access the twitter data, we authorize our application by using the following code snippet in R:

```
setup_twitter_oauth(CONSUMER_KEY,
                    CONSUMER_SECRET,
                    OAUTH_TOKEN,
                    OAUTH_TOKEN_SECRET)
```

### 5.2.4 Creating Data Frames

Now we extract the tweets using GET search/tweets source for a specified query. Data selection based on the data structure is an important role for data analysis. Data organized into the tables gives the better view of the data. For the tweet analysis we use the Data Frames enabled with R. Data Frame is capable of storing the data in the table form and label its rows and columns. Fig 5.6 shows the tweets in Data Frame.

| | text | favorited | favoriteCount | replyToSN | created | trunc |
|---|---|---|---|---|---|---|
| 1 | RT @Madan_Chikna: People from Mumbai expressing... | FALSE | 0 | NA | 2017-04-26 08:33:02 | FALSE |
| 2 | RT @Abhina_Prakash: That Delhi chose Dengue &am... | FALSE | 0 | NA | 2017-04-26 08:33:02 | FALSE |
| 3 | RT @smritiirani: Congratulations to karyakartas &a... | FALSE | 0 | NA | 2017-04-26 08:33:02 | FALSE |
| 4 | RT @RepublicofIndia: #Sensex Hits Record High 301... | FALSE | 0 | NA | 2017-04-26 08:33:01 | FALSE |
| 5 | RT @vivekshettym: Once again !#MCDresults #MCDE... | FALSE | 0 | NA | 2017-04-26 08:32:57 | FALSE |
| 6 | RT @MODIfyingBHARAT: Thanks for the trust shown ... | FALSE | 0 | NA | 2017-04-26 08:32:57 | FALSE |
| 7 | On a day when his party has won a landslide victory ... | FALSE | 0 | NA | 2017-04-26 08:32:55 | TRUE |
| 8 | RT @himantabiswa: Trends in #MCDresults yet again... | FALSE | 0 | NA | 2017-04-26 08:32:54 | FALSE |
| 9 | RT @Republicofindia: BJP winning in Jama masjid , is ... | FALSE | 0 | NA | 2017-04-26 08:32:54 | FALSE |

**Fig. 5.6: Tweets Data Frame**

### 5.2.5 Data Preprocessing

The extracted tweets contain the noisy data because tweets are written in an informal language which has no structure. By specific domain, it is difficult task to categorize tweets. So clean the data for the analysis of twitter data which is a difficult job. To start with retweets, remove the duplicate tweets because they already convey the same information many times. Then remove the stop words such as is, are, to, the, more etc. Stop-words don't have any value for sentiment analysis, as there is no context in the meaning. Then eliminate the web references or web links. Also remove the special character or symbols, white spaces, punctuation marks which includes comma, semi colon, Question Mark, Underscore etc. Then remove the numerical digits. Then transform each tweet into the lower case format. After that, using the classical tokenization method split the raw data into the separated words separate by space, comma, special character etc. The process of splitting a block of the text into the tokens which is represented by words in the sentence is known as Tokenization. Finally the clean data will be delivered for the analysis.

### 5.2.6 Frequency Analysis

Various analysis methods can be applied on the twitter data which is extracted in data frame. Frequency Analysis is defined as to count the frequency of the words. It tells us that which word is discussed more by the users. In order to explore the frequency of words we can apply the *TermDocumentMatrix()* from text mining. The frequency of words contained in the table which is called Document Matrix. Row names are documents and column names are words.

### 5.2.7 Clustering of Twitter Data

The clustering algorithm is to be applied on the fetched data based on the information of location. The tweets which are unique and contain only the geographical information are considered for the clustering. The tweets with geographical information means we consider only those tweets which contains latitude and longitude information. The k-means clustering algorithm is applied on the data to classify the twitter tweets into the geographic clusters.

### 5.2.8 Visualizing Geographic Clusters with Google Map

For analysis of social data, visualization plays a vital role. The huge amount of data can be interpreted by using the visualization method but when it is visualized in a bad way, it represented the distorted information. K-Means Clustering algorithm is to be used to visualize and cluster the tweets by plotting it on Google Map. The location information is already fetched through the Twitter API that describes the latitude and longitude coordinates of the location. Then we plot the map in a tool like Google Earth. The Elbow method in R can calculate the number of clusters and then apply the K-Means to group the tweets by location, cluster them and then visualize on the Map using the *ggmap* package in R.

## 5.3 Results and Discussions

One way to analyze the data is through data visualization. Data visualization is the representation of data in a graphical or pictorial format. It helps the decision makers to recognize new patterns or grasp difficult concepts to see analytics visually. Using technology we can take the concept ahead with interactive visualization in order to make charts and graphs leading to move detailed description of how we visualize and how the data is processed. Word Clouds, maps, bar chart, pie chart, histogram etc. can be used for visualization.

### 5.3.1 Frequency Analysis of Words

We are taken the hashtag “MCDResults” which is based on the MCD election on Delhi 2017. As discussed earlier, twitter data is extracted using Twitter API and then we preprocess the data and now data is cleaned from the stop words, URLs, punctuation marks etc. Then the frequency analysis is to be done on the cleaned data.

The frequency of words is shown in Fig 5.7 in decreasing order. “aap” word is discussed more by the users with the frequency 715.

| | word | freq |
|---|---|---|
| aap | aap | 715 |
| delhi | delhi | 510 |
| dengue | dengue | 358 |
| bjp | bjp | 297 |
| amp | amp | 260 |
| kejriwal | kejriwal | 253 |
| one | one | 200 |
| people | people | 189 |
| todays | todays | 183 |
| meanwhile | meanwhile | 180 |
| meeting | meeting | 178 |
| mybit | mybit | 175 |
| time | time | 169 |

**Fig. 5.7: Top words and their frequencies**

The word used in the given hashtag “MCDResults” is shown in a bar plot according to the frequency of the words. As shown in Fig 5.8 “aap” word is used more regarding search term “MCDResults”. The x-axis depicts the words and y-axis depicts the frequency of words.

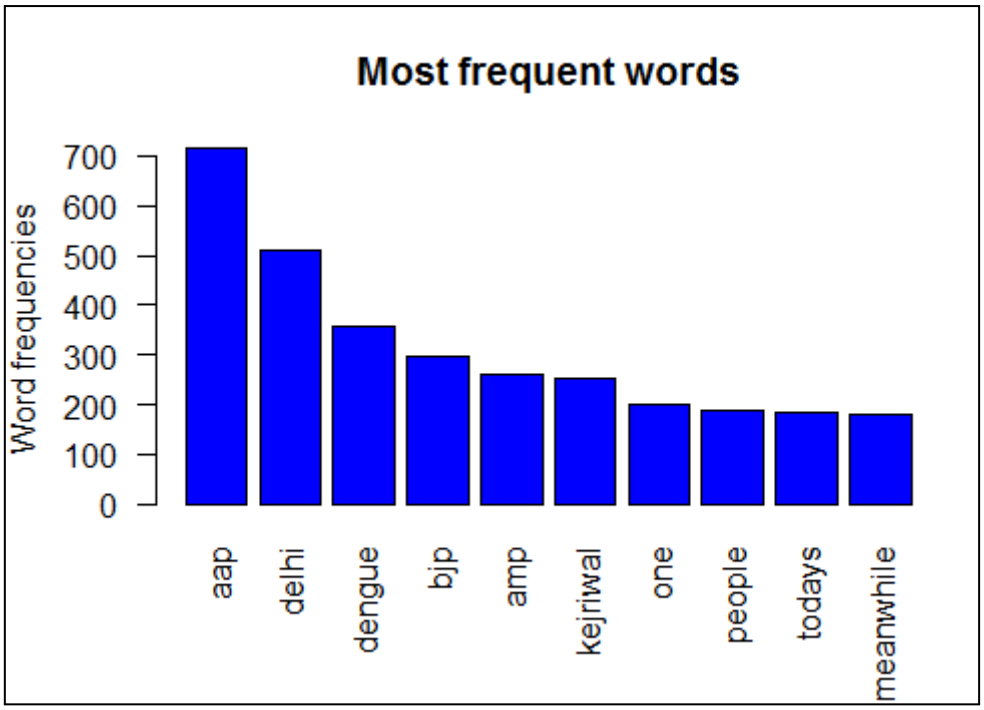


**Fig. 5.8: Histogram of Top word with their frequencies**

To visualize free form text and to understand the keyword metadata (tags) on websites, a word cloud is graphical represented for text data. It is pictorial

representation of word frequency. The *tm* and *wordcloud* package is used in R to represent the most used words associated with the hashtag in a pictorial representation.

#MCDResults is a very popular hash-tag in 2017 because of MCD Election Results in Delhi. Delhi, BJP, political affairs, and election controversy have been broadly discussed using this hashtag. For political discussion, this hashtag became the most popular online platform. To understand the gathered tweets better, the researcher have used word cloud. Fig 5.9 shows the word cloud for word "MCDResults".

Firstly, it counts all the words from the tweets of the given hashtag and then ranks it. Then the rank is analyzed and presents graphically as their count size. The size of the word depends upon the count of the word. In the Fig 5.9, the frequency of the word "aap" is the highest. This parameter gives the broad idea of the internal content ranking through special graphical visualization.

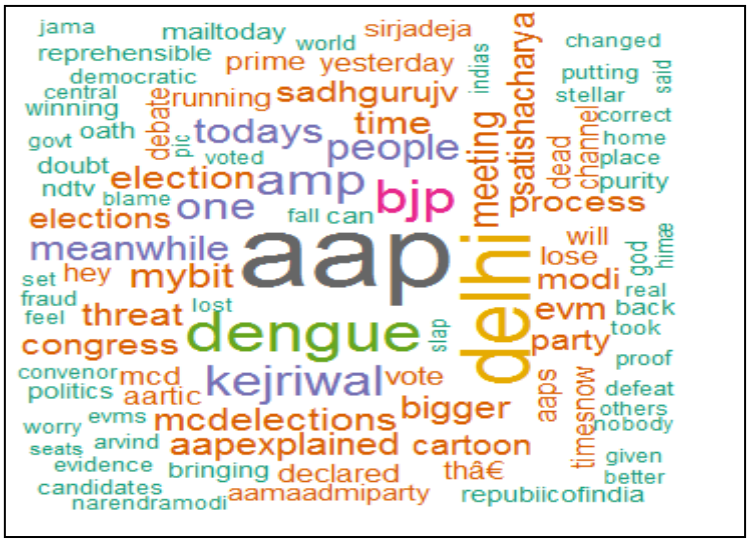


**Fig. 5.9: The Word Cloud**

### 5.3.2 Geoclustering of Twitter Data

We first extracted the twitter data in the data frame format from the #MCResults hash tags using the *Twitter Graph API* in R. Then we have considered the unique tweets which contains the geo tagged information using the *Google Map Geocoding API* in the *ggmap* package in R. Fig 5.10 shows the spatial distribution of tweets. Every dot represents a single tweet and we restricted the area to India.

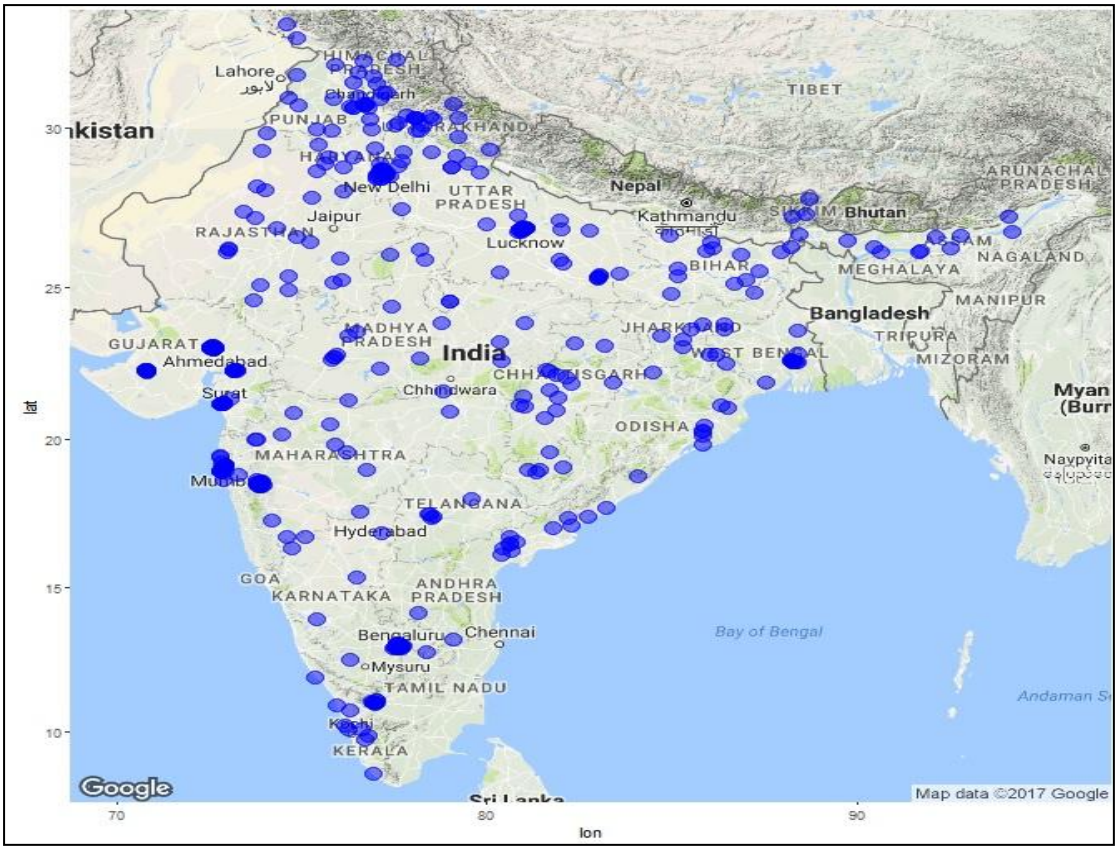


**Fig. 5.10: Location of each tweet**

After that the k-means clustering algorithm is applied to cluster the tweets. The number of clusters is predefined and selected randomly or generated from the data set using *Elbow method* in the k-means clustering algorithm. The idea behind the elbow method is to apply the k-means clustering algorithm on the dataset for a range of values and calculate the Sum of Squared Error (SSE) for each value. Then plot a line chart for each value of the SSE. If the chart looks like an arm, then the value of the elbow of the arm is the best. The idea is that we want a small SSE, but as we increase the value, the SSE tends to 0. So choose a small value that still has a low SSE. Fig 5.11 shows the number of clusters found using *Elbow Method.*

The above figure calculates the number of cluster required for the k-means clustering algorithm using the Elbow Method. The line chart is between the Sum of Squared Error (SSE) and the number of clusters. Here the elbow value indicates the number of clusters which are best suitable for our twitter dataset. The above figure indicates that the value of elbow is 6 which is the number of clusters for our dataset.

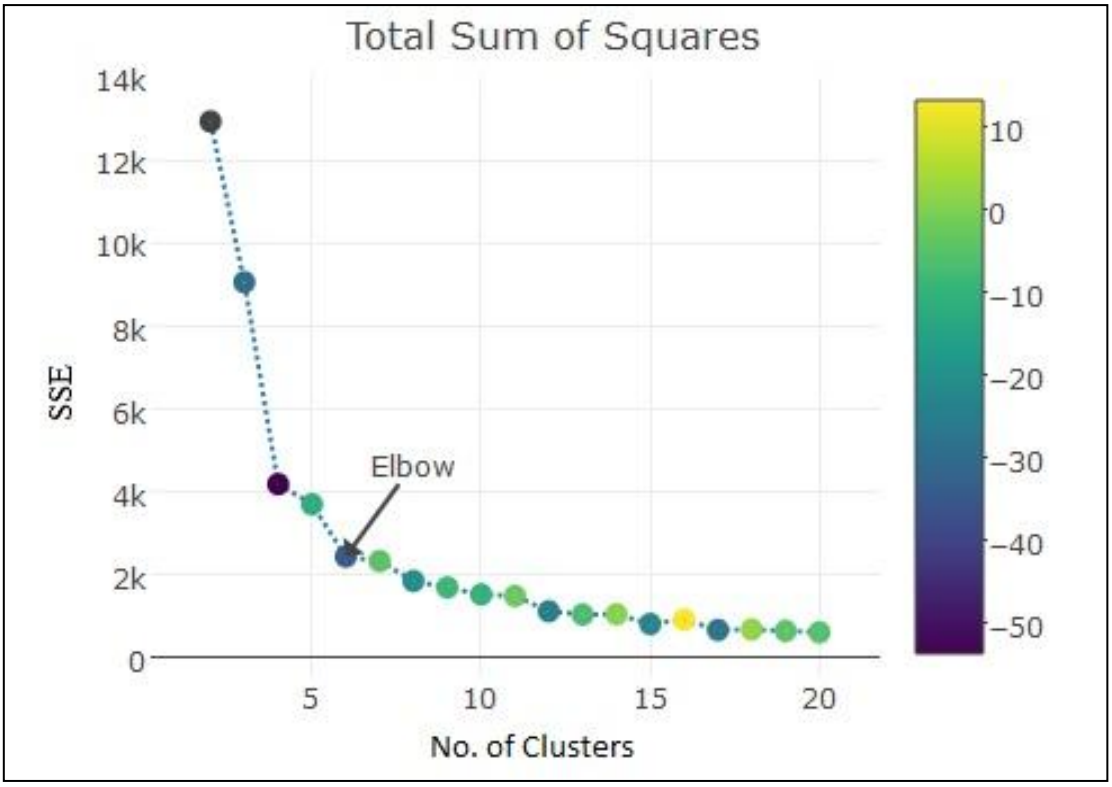


**Fig. 5.11: Finding number of clusters using Elbow Method**

After calculating the number of clusters required for the dataset using *Elbow method*, the k-means clustering algorithm is applied based on the geotagged information. The result of the clustering algorithm is represented in Fig 6. In the figure below, it can be clearly observed that in the northern parts of India there is maximum number of tweets regarding the Delhi MCD results. Specifically, Delhi itself has the maximum number of tweets in northern part of India as shown in the Fig 5.12. On the other hand, when compared to northern parts of India, the western parts have comparatively less tweets. The number of tweets in central, southern and eastern parts of India is similar in number.

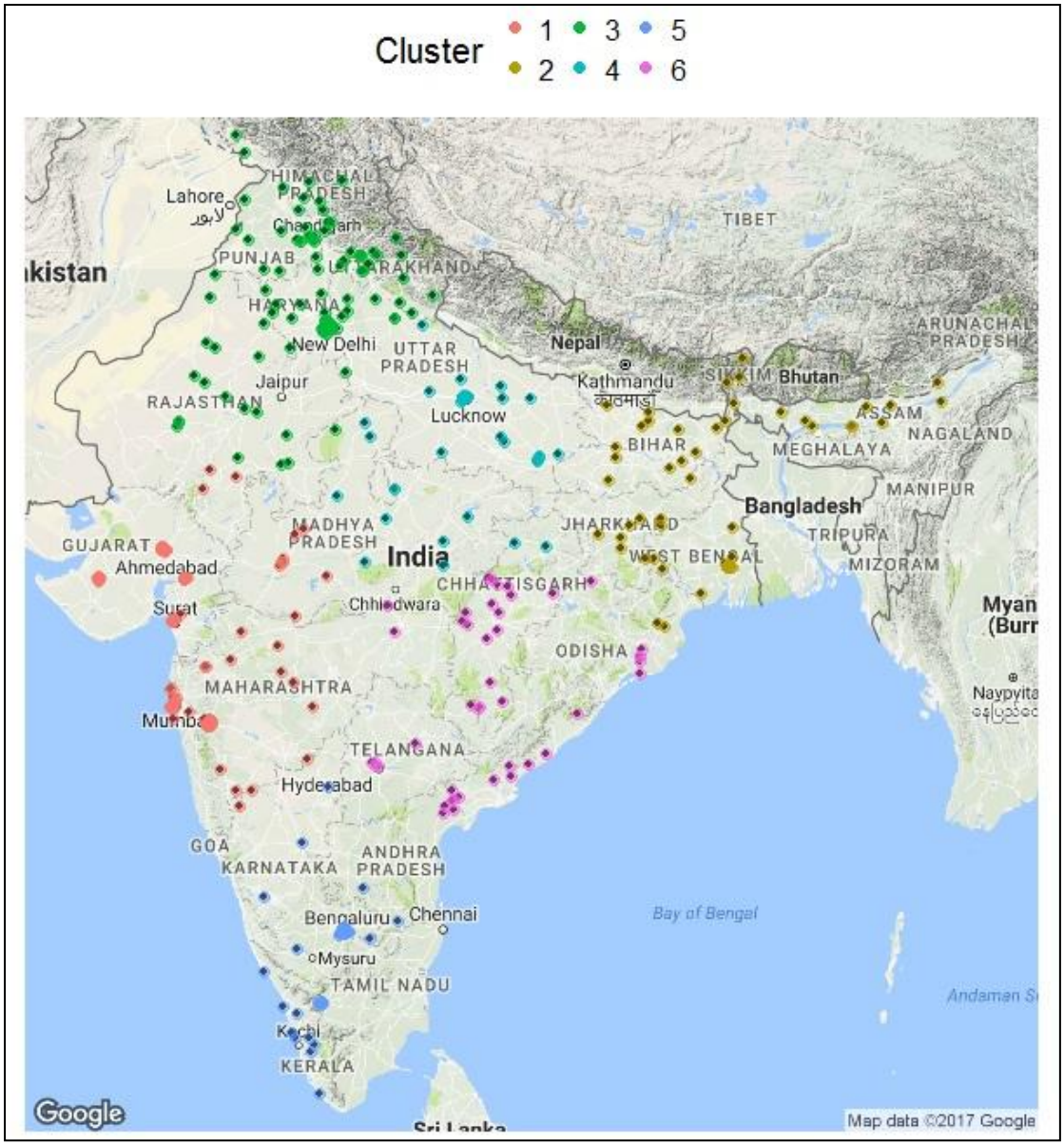


**Fig. 5.12: Clustering the tweets**

Fig 5.13 further shows the overlay of tweets represented in Fig 6. *MASS* package and *2-D Kernel density estimation* function have been used to estimate the overlay onto plot. In case there are new tweets with the help of overlay plot we can classify them into specified clusters.

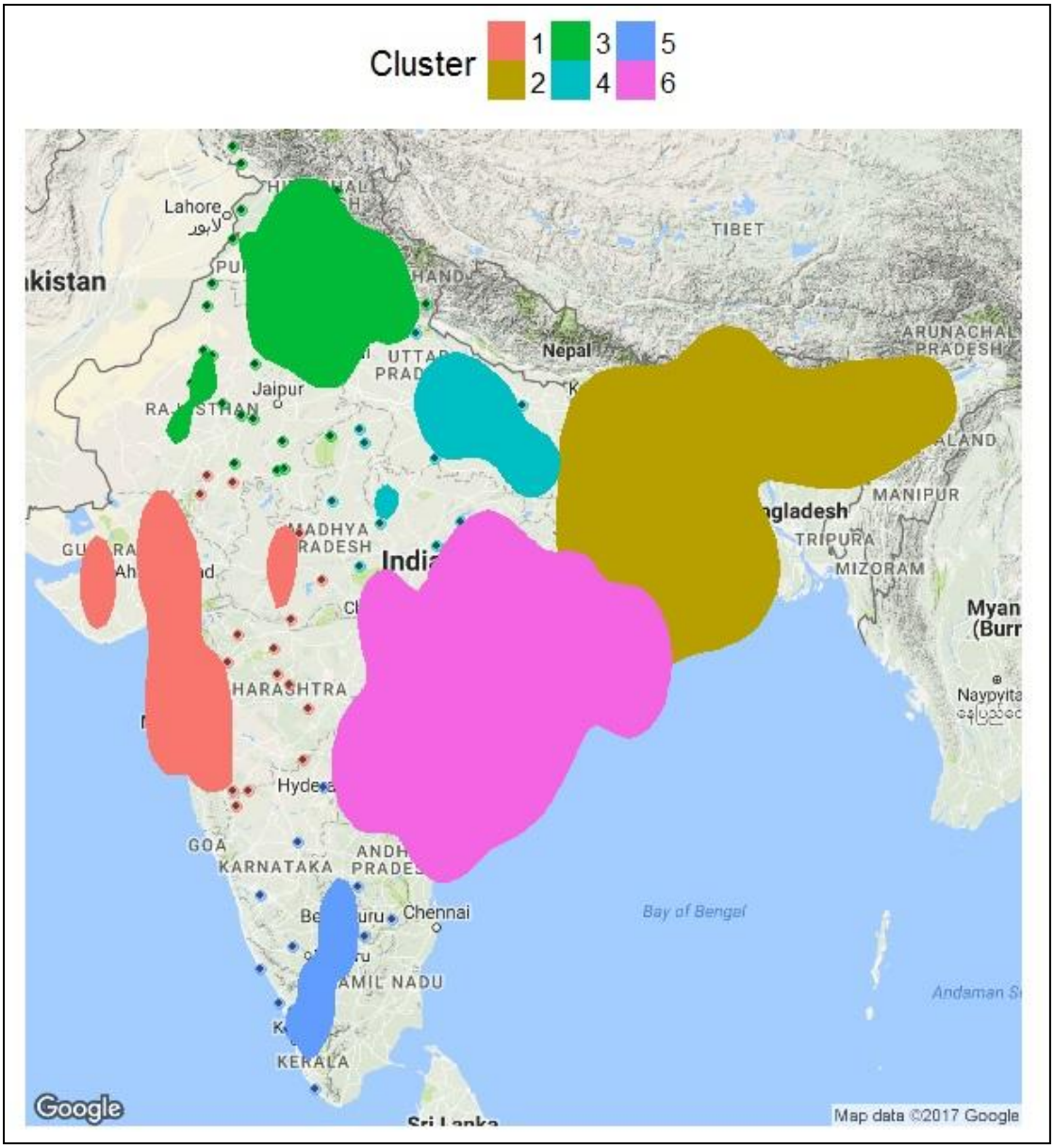


**Fig. 5.13: Overlay visualization of Clusters**

# CHAPTER 6 ANALYSIS AND VISUALIZATION OF LINKEDIN DATA

The business and professional relationship are focused by the powerful and popular social network site which is LinkedIn. LinkedIn is launched in 2003. It provides a platform to the individual so that they post their resume, search a job according to their education and professional qualification, connect with other individual of same professional qualification and recommend their friends. LinkedIn may appear like other social networking sites but the data provided by LinkedIn is different as the data provided by the other social networking sites like Facebook, Twitter etc. Currently, LinkedIn has 500 million members from 200 countries [38]. The site is available in 24 languages [39].

The data available at LinkedIn is very unique in relation to the data available at the other social networking sites because of the professional characteristics of LinkedIn. LinkedIn provides the large amount of data of the millions of the subscribers. This large amount of dataset is not able to handle or manage by the traditional database management techniques. The dataset become complex due to the number of subscribers has increase in multiple of thousands [13]. The dataset is very simple if it contains less than 2000 profiles. The researchers have proposed various tools and techniques to store compute and handle the large amount of dataset.

To manipulate these complex and huge datasets is a difficult task. This can be possible by dividing these datasets into small subsets by using the specific tools and methods for constructive outcomes. A lot of information can be collected about individuals and groups by investigating the LinkedIn data. The LinkedIn provides the information (school and college information, employer, profession, past history of job etc.) of the individuals to the users even they are not connected to the network [40]. The semi structured data is provided by the LinkedIn members. LinkedIn members freely give away job titles, industry information, location information, skill sets, educational qualification etc.

The behavior and properties of individuals and group can be understood by investigating the LinkedIn data. The professional relations of individuals and groups

can be explored by analysis the LinkedIn data. The personality of the individual can be picturized by collecting all the LinkedIn relations, the previous background of the designation, occupation, locality, employer etc. Some parameters can be set for LinkedIn in order to compare individuals. Various types of professional communities can be detected by hierarchical clustering on the basis of job title and companies by analyzing the LinkedIn dataset.

## 6.1 Implementation Methodology

The implementation methodology is shown in Fig 6.1. In the initial step, the data is obtained by exporting the LinkedIn connection as an address book. In the second step, normalization of LinkedIn data is done to make the data valuable for further analysis. Then the frequency analysis and similarity measurement of company names and job titles are carried out. Then apply the hierarchical clustering algorithm on the dataset as per the distinctive similarity criteria such as company names and job titles. Finally the results of the obtained clusters are visualized for analysis.

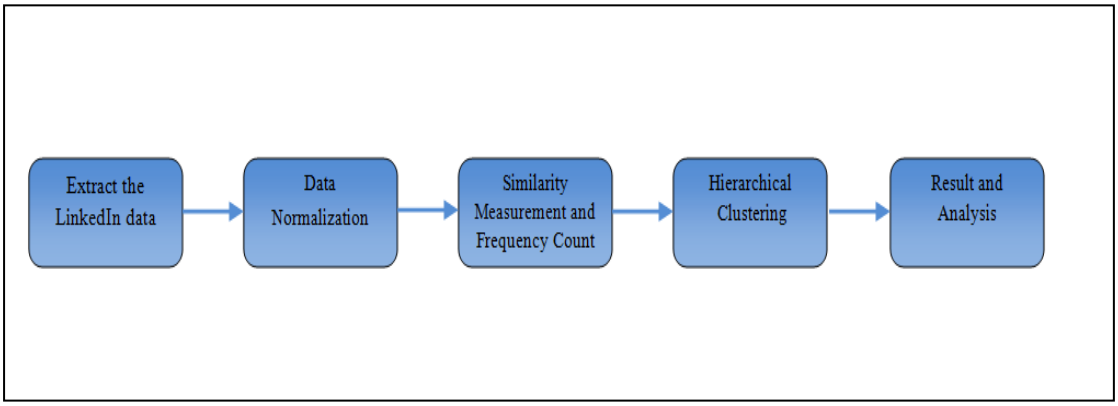


**Fig. 6.1: Methodology of Proposed Model**

The various packages are available in python to analyze and visualize the LinkedIn data. The python packages can be downloaded and installed using *pip*. Table 6.1 describe the purpose of various packages used for analysis and visualization of twitter data.

**Table 6.1: Various Packages in Python for Analysis and Visualization**

| Python Packages | Description |
|---|---|
| PrettyTable | Displaying the tabular data in the ASCII table format |
| Counter | Subclass for counting the hashable objects |
| Csv | Read and write the tabular data in CSV format |
| Cluster | Implements various clustering algorithm and creating several clusters of object from a list |
| Nltk | Used for natural language processing |
| Webbrowser | Provides the interface to display web based documents to user |
| Shutil | Number of operations on files and collection of files |

## 6.2 Data Extraction

Earlier, the LinkedIn provides a service in which user obtains their profiles and networks connected to them via OAuth based Application Programming Interface (API). To retrieve the data using this API, there was need to register a new application on *https://developer.linkedin.com/* with the LinkedIn. The authentication keys were generated which includes Client ID, Client Secret, OAuth Key, OAuth Secret which was used for the authorization of your application and the data is generated in JSON or CSV format. But in 2015, the professional network LinkedIn was restricted access to most of its API [43]. So we are not able to retrieve the data from the API. LinkedIn allows us to export our connections to a CSV file. By exporting the connections from LinkedIn, first log into the LinkedIn. Then click on the 'My Networks' tab and then click on the 'Managed synced and imported contacts' which is on the top right corner of the page. In *advanced actions* section, click on the 'Export contacts'. In the next page, we have two options to choose the format that we prefer. The formats are 'Fast file only' and Fast file plus other data. The Fast file only includes the basic information like profile information, connections and messages while other includes the detailed information like account activity and history. By selecting any of format, click on the Request archive. We will receive an email from the LinkedIn with a link. By clicking on the link that we received from email, the page is opened which is shown in Fig 6.2. Click the download button to retrieve the data in CSV format.

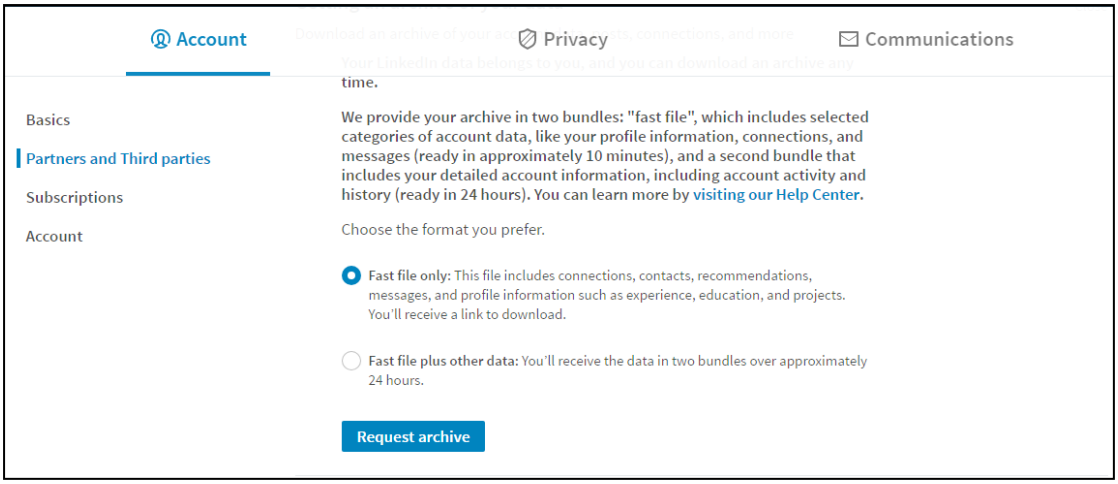


**Fig. 6.2: Archive the Data**

We can also directly retrieve the data by exporting the LinkedIn connection as an address book from the *https://www.linkedin.com/people/export-settings* which is shown Fig 6.3. There are various options to export the data in different formats. The different formats are selected from 'export to' tab. The formats are Microsoft Outlook (.CSV file), Outlook Express, Yahoo! Mail, Mac OS X Address Book and vCard. For our analysis we select, Microsoft Outlook (.CSV file) format.

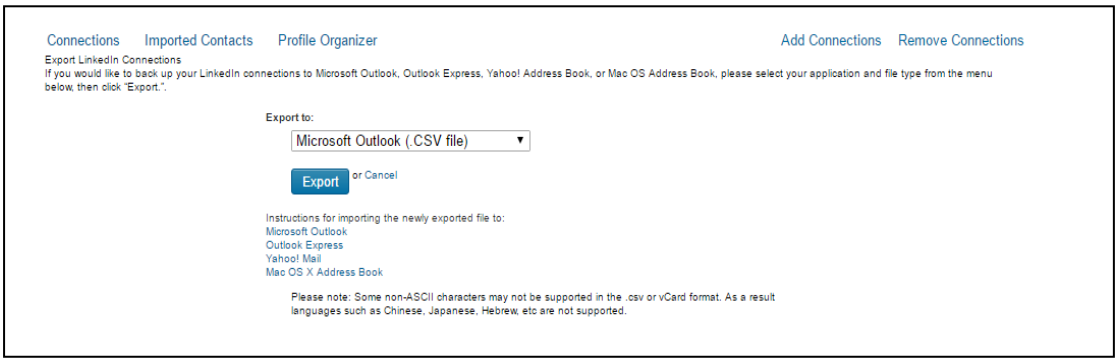


**Fig. 6.3: Export the Connections**

### 6.2.1 Extracted Dataset and its Features

The dataset is obtained by exporting the LinkedIn connection as an address book in CSV format. The description of the various attributes that can be obtained by exporting the data set from the LinkedIn connection as an address book is shown in Table 6.2.

**Table 6.2: Features of the Dataset**

| Attributes | Description |
|---|---|
| First Name | First name of person |
| Last Name | Last name of the person |
| Email Id | Email address of the person |
| Company | Name of the Company where the person works |
| Job Title | Describes the position of the person |
| Location | The place where the person works |

### 6.2.2 Clustering the LinkedIn Connections

Clustering, an unsupervised machine-learning technique is used as a fastener in all the data mining toolkits. It is defined as collection of objects in one group which are similar between them and dissimilar to other group. It is extensively used in various fields such as text mining, image analysis, machine learning, image processing, web cluster engines, weather report analysis, bioinformatics, etc [41]. The main step for clustering analysis of LinkedIn or any other social network is Data Normalization. LinkedIn users are able to enter the free text in their professional information which result the inconsistency of data like misspelled, abbreviations, blank columns etc. For the analysis of data in an efficient manner, firstly normalize the data and then use the efficient clustering technique.

When data is obtained by exporting the LinkedIn connection as an address book, the data is not in the exactly same format as we had liked because members can enter the data related to job titles, location and company in various ways. First we tokenize the data. Tokenization is defined as slicing a block of text or content into a set of words or token. This includes splitting the text using a delimiter like white spaces, comma, slash, etc., expanding words, expanding abbreviation, etc. Normalize the data when the tokenization is done. The normalization of data includes having consistent token over a similar set of words [42]. The tokenization and normalization of data can be achieved in python using nltk package [44]. Thus, if we want to gather all the members of "TCS" as their organization, at that point our search ought constrained to "TCS", as well as "Tata Consultancy Services" or "TCSL" or "Tata Consultancy Services Ltd." or "Tata Consultancy Services Limited" and many more in our search domain. A lot of effort is required to get the data into a form that is suitable for analysis. Redundancies present in various features like job title name, company names

etc can be removed by normalizing the data. The aggregate titles are segregated using slash like organizer/convener and replaced with popular short forms.

## 6.3 Results and Discussions

Firstly, the fetched data stored by exporting the LinkedIn connection into the address in a CSV file is to be pre-processed for the further analysis. A lot of effort is required to get the data into a form that is suitable for analysis. Redundancies present in various features like job title name, company names etc can be removed by normalizing the data. The aggregate titles are segregated using slash like organizer/convener and substituted with popular short forms. The *prettytable* package in python is used to display the output in a tabular format. Results are obtained after processing data through frequency analysis is shown in Fig 6.4

| Company | Frequency |
|---|---|
| Thapar University | 14 |
| Tata Consultancy Services | 4 |
| Oracle | 3 |
| IIT | 2 |
| Infosys | 2 |

**Fig. 6.4: Frequency count of Company name**

The LinkedIn data can be exported as address book in a CSV file followed by a standardization and exhibition of dataset in a primary manner. This is implemented by selecting a suitable set of information and breaking it down into subsets. For example, for searching the name of company “Tata Consultancy Services”, we have to strip off general suffix like “LLP”, “INC”, “LLC” and many more.

In this phase the data is analyze as per the job title. The data is inconsistent because of the different modes of input by various users. The data needs to be managed into the number of subsets. Job title like the company name can also be arranged in a specific manner. The need for normalization arises due to same information conveyed by ambiguous job title like some user input the job title “C.E.O” and some “Chief Executive Officer” which shows the same job title. A specific solution in these regards would be a normalization followed by appropriate classification for review

purpose. A similar approach is applied in this study which included a normalization followed by frequency analysis which transforms the job title into suitable clusters. The data has been normalized according to the job titles and frequency analysis of the normalized data is shown in Fig 6.5 The aggregate titles are segregated using slash like organizer/convener and substituted with popular short forms. Results are obtained after processing data through frequency analysis is shown in Fig 6.5. The frequency count of various tokens present in the job titles are shown in Fig 6.6.

| Job Title | Frequency |
|---|---|
| Student | 12 |
| Research Scholar | 2 |
| Internship Trainee | 2 |
| Project Intern | 2 |
| System Engineer | 2 |

**Fig. 6.5: Frequency count of Job Title**

| Token | Frequecny |
|---|---|
| Student | 13 |
| Engineer | 7 |
| System | 4 |
| Project | 3 |
| Senior | 3 |
| Research | 2 |
| Specialist | 2 |
| Internship | 2 |
| Scholar | 2 |
| Consultant | 2 |
| Software | 2 |
| Intern | 2 |
| Trainee | 2 |
| Developer | 2 |

**Fig. 6.6: Frequency count of token in Job Title**

### 6.3.1 Clustering Job Titles and Visualization of obtained Clusters

The clustering is to be done based on the job title and company name by finding the similarity between them by using Jaccard distance similarity. The Hierarchical Clustering module is to be import which comes under the cluster package in python for hierarchical clustering. For similarity between the clusters, we need to import the jaccard_distance module that comes under the nltk package. Fig 6.7 displays the hierarchical clustering based on the job titles and shows the connection accordingly. In order to manage and filter data in subparts according to analyzing criterion this approach can be followed. A significant label is given to each cluster. We compute significant labels by taking the set wise intersection of various terms present in the job titles for every cluster. The user who shares the roles in the job duties are in the same cluster and this information is useful in this case.

```
Descriptive Terms: Internship, Trainee
--------------------------------------
Niyati Trivedi
Sonia Mittal

Descriptive Terms: System, Engineer
-------------------------------------------
Navneet Kaur
Radha Shukla
Tuhina Mehta

Descriptive Terms: Project, Intern
----------------------------------
Kanika Sharma
Niharika Verma

Descriptive Terms: Software, Developer
------------------------------------------
Gaurav Patel
Ginni Rani

Descriptive Terms: Student
--------------------------
Abhishek Kapoor
Ashish jat
Guneet Kukreja
Jagmeet Kaur
Kapil Jindal
```

**Fig. 6.7: Clustering Job Titles using Hierarchical Clustering**

After that, the above result of the clustering of our connections can be representing in two graphs one is node link layout and another is dendrogram using D3.js (the state of art visualization toolkit). The leaves of the tree represents the individual people, on the other hand the nodes represents the information such as “Project Intern” combine the leaf nodes to form a cluster. The node link tree and dendrogram is shown in Fig 6.8 and Fig 6.9 respectively. A large amount of data is clearly visible when we observe the simple images of the professional network.

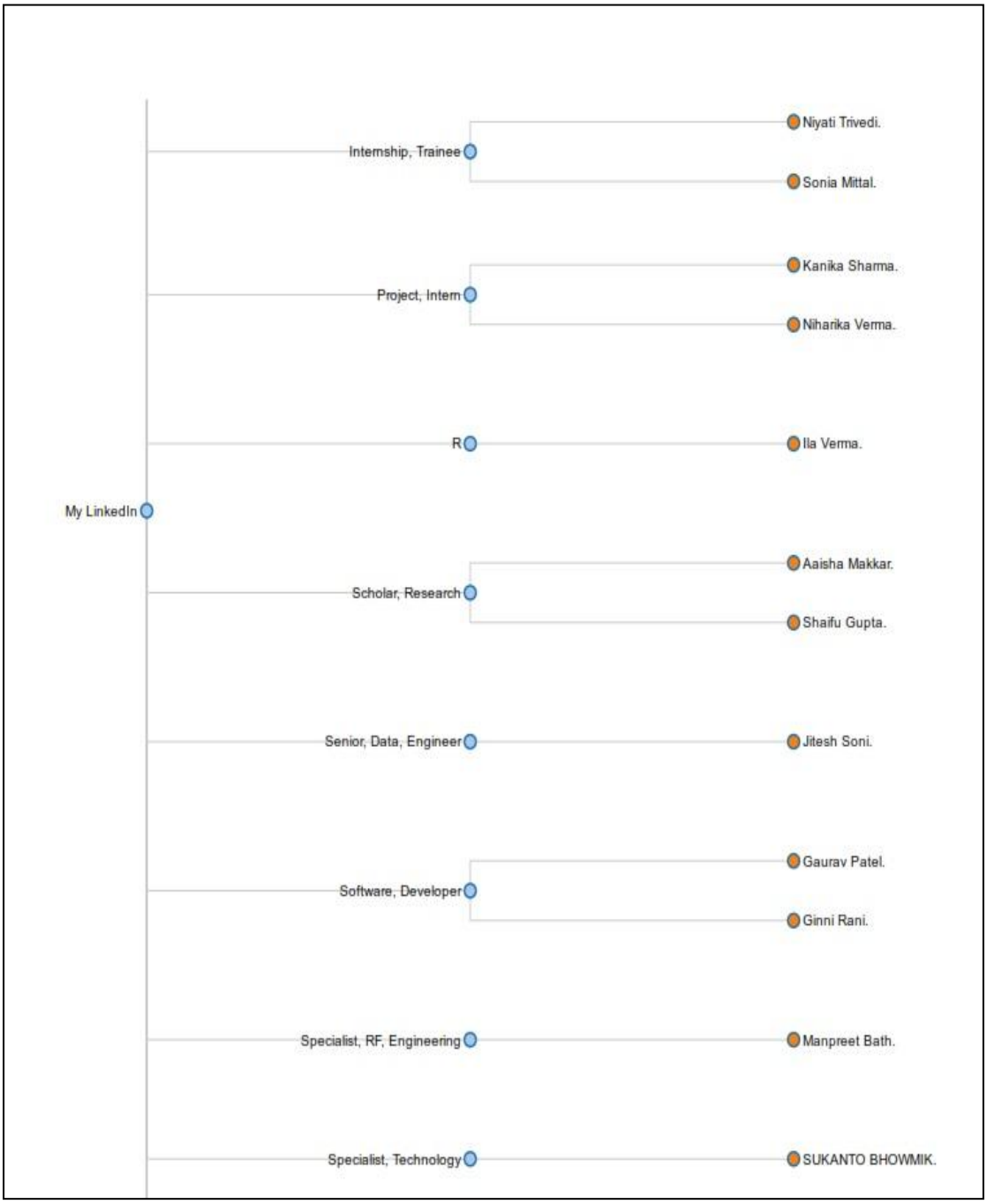


**Fig. 6.8: Dendrogram Layout of Contacts Clustered by Job Title**

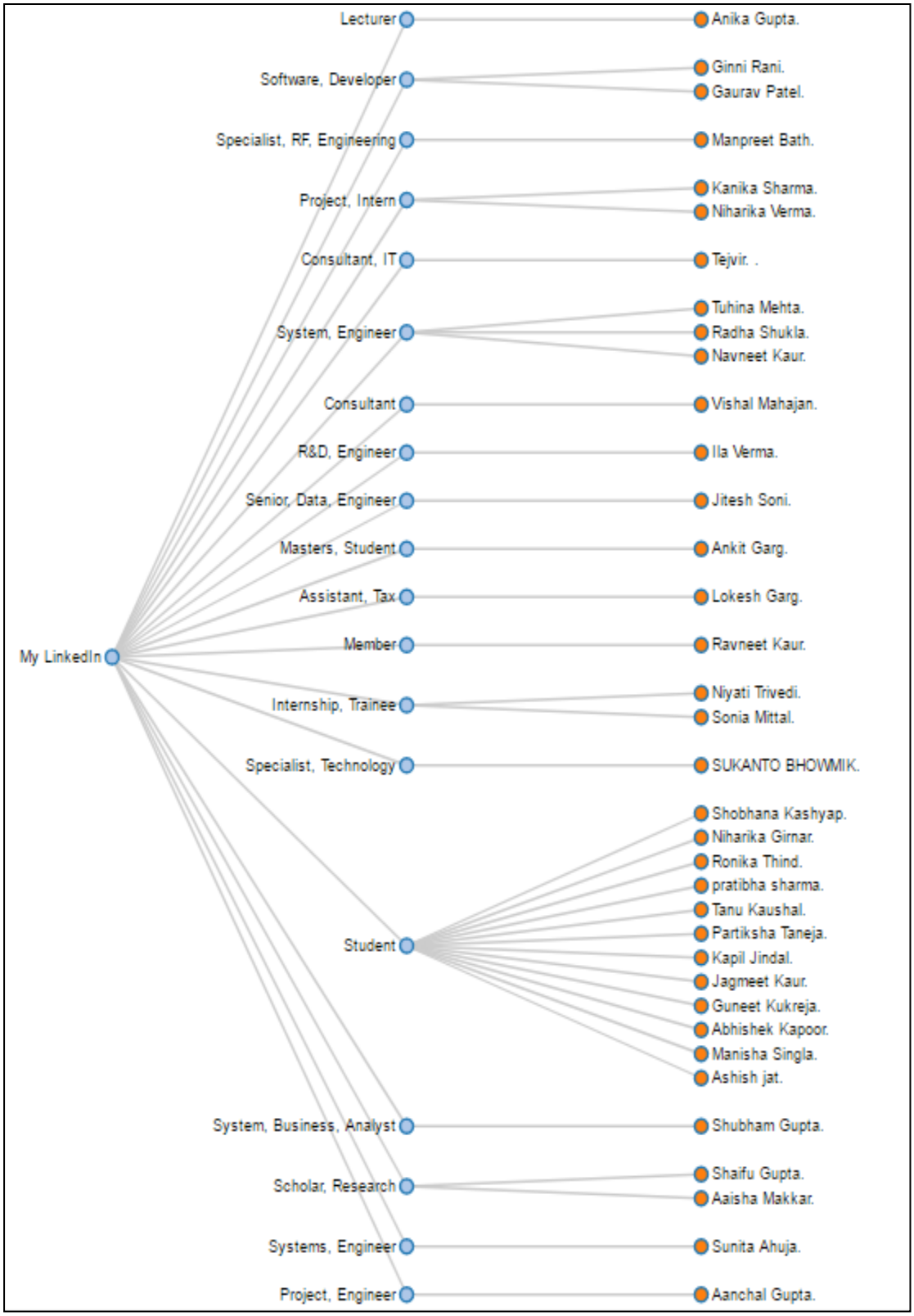


**Fig. 6.9: Node-Link Tree Layout of Contacts Clustered by Job Title**

### 6.3.2 Role of Clustering in Enhancing User Experience

The clustering can group the individuals together who are likely share the common responsibilities based on the job title. The powerful tool to visualize the clustered data in hierarchical form is the Dojo Tree component which is simple and efficient. The result of clustered data based on the job title in tree widget form is shown in Fig 6.10. This approach provides a controlling option for filter the search criteria. It also displays the number of connection in each group. This type of information might be useful for variety of reasons like to call the Research Scholars for a conference.

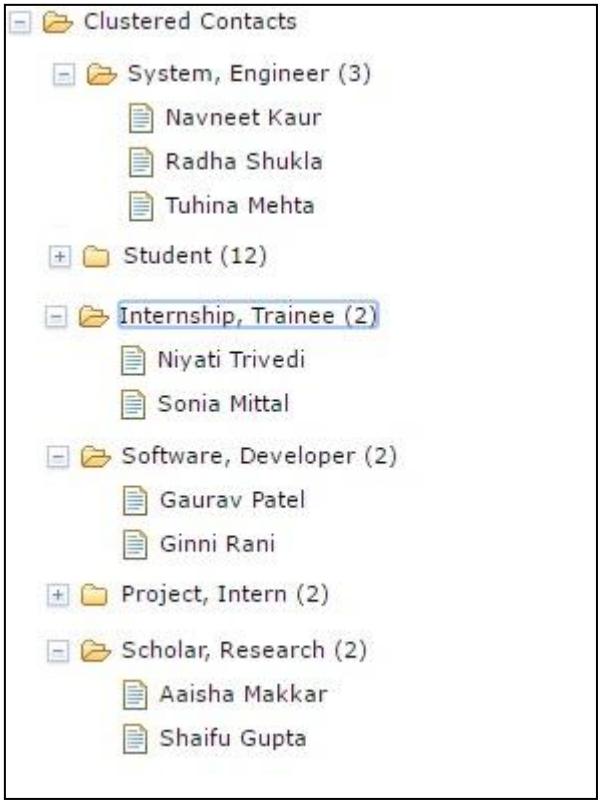


**Fig. 6.10: Displaying Intelligently Clustered Data enhances User's Experience**

# CHAPTER 7 CONCLUSION AND FUTURE SCOPE

## 7.1 Conclusion

In this thesis, various approaches have been described which are used to detect the communities and have been compared based on the modularity using the real world networks and also analyze the social web to explore the research issues associated with it. Afterward Twitter and LinkedIn the two famous social networking sites have been used as a data source for the purpose of community detection, frequency analysis and geo-clustering. The fore mention process concludes the analysis of visualization of data.

We reviewed the various non-overlapping and overlapping community detection algorithms such as Graph Partitioning, Label Propagation Algorithm, Hierarchical Clustering, Girvan Newman Algorithm, Partitional Clustering, Clique Percolation Method etc for both the directed and undirected networks. Some of the existing algorithms are only suitable for the small scale networks not for the large scale networks. Some of these algorithms are examined on the different datasets. We downloaded the massive social dataset which is available on the various websites and utilize them to detect the communities for different algorithms and compared the algorithms. The major issues identified through the thorough analysis of algorithm comprises of the scalability and the quality of the detected communities.

Secondly, in the research we analyze and visualize the twitter data. For this analysis, we extract the tweets from the twitter about the "MCDResults" by using twitter APIs. The language used to extract the tweets is R. R language is used for the data acquisition, preprocessing and visualization of the clustered data. As R is a popular language used for retrieving, cleaning, analyzing and visualizing the data. Therefore, features of R are utilized for sentimental analysis and visualization. The preprocessing task is needed to analyse the tweets and viewed in the data frame format. The analysis of extracted tweets is performed by the frequency analysis. Web 2.0 introduced the concept of word cloud, which helps to visualize the content of a website, the larger the words appeared on a widget, and the more frequently they appeared on the site.

The location information present in the Twitter dataset is also utilized in geo-clustering. For the geo-clustering, we considered the tweets which are unique and contain the latitude and longitude information of the location. After that, the k-means clustering algorithm is applied to cluster the tweets. These clustered tweets are then visualized on the Map using the *ggmap* package in R.

Thirdly, in the research we analyze and visualize the LinkedIn data. For this analysis, we have exported the LinkedIn data from the LinkedIn connection and stored in the CSV file. The language used for the analysis of LinkedIn data is Python. Python is a powerful language for building data analytics tool. So we utilized the various packages of python for analysis of clustered data. To visualize the clustered data in tree and dendrogram we use the D3.js visualization toolkit. We cluster the LinkedIn data using the hierarchical clustering technique based on the job title and the company name and also overcome with few common problems like normalization of the inconsistent data. The useful information of our professional LinkedIn contacts can be analyzed and visualized by mining the LinkedIn data which cannot be discovered manually.

## 7.2 Future Scope

A strategically reduction in the computational time could be achieved by modifying the clustering methodology or by taking the advantage of high performance computing framework like Hadoop, Graphical Processing Unit (GPU).

Although the variety of pre-processing techniques exist and could be explored for higher precision by an accurate feature identification. But due to the time constraint only basic feature extraction technique have been explored for this study. Also there is a scope for extension via coupling with natural language processing.

By analyzing the LinkedIn data, the preferred jobs can be recommended to our contacts based on the job skills and past experience by designing the predictive model.

By connecting LinkedIn API and twitter API, the various trending topics can be fetched in different professional communities.

Such a linking would enable us to classify the twitter tweets from a professional perspective or otherwise and thus introduce scope for sentimental analysis.

Ultimately sentimental analysis has lot of scope for improvement and development along with a promising future in a broader field of research.

# REFERENCES


[1] R. Baeza-Yates, "Graphs from search engine queries," in *International Conference on Current Trends in Theory and Practice of Computer Science*. Springer, 2007, pp. 1–8.

[2] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the web for emerging cyber-communities," *Computer networks*, vol. 31, no. 11, pp. 1481–1493, 1999.

[3] A. Costill. (2013, Dec 10). 25 Insane Social Media Facts [Online]. Available: http://www.searchenginejournal.com/25-insane-social-media-facts/79645/

[4] M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of social media," *Business horizons*, vol. 53, no. 1, pp. 59–68, 2010.

[5] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? get serious! understanding the functional building blocks of social media," *Business horizons*, vol. 54, no. 3, pp. 241–251, 2011.

[6] [Online]. Available: https://womseo.com/wp-content/uploads/2012/12/The-Social-Media-Marketing-Report-2012.pdf.

[7] G. Palla, I. Der´enyi, 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.

[8] F. Luo, J. Z. Wang, and E. Promislow, "Exploring local community structures in large networks," *Web Intelligence and Agent Systems: An International Journal*, vol. 6, no. 4, pp. 387–400, 2008.

[9] S. Gregory, "Finding overlapping communities in networks by label propagation," *New Journal of Physics*, vol. 12, no. 10, p. 103018, 2010.

[10] J. Chen, O. Za¨ıane, and R. Goebel, "A visual data mining approach to find overlapping communities in networks," in *Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in*. IEEE, 2009, pp. 338–343.

[11] E. R. Barnes, "An algorithm for partitioning the nodes of a graph," *SIAM Journal on Algebraic Discrete Methods*, vol. 3, no. 4, pp.541–550, 1982.

[12] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *The Bell system technical journal*, vol. 49, no. 2, pp. 291–307, 1970.

[13] P. Garg, R. Rani, and S. Miglani, "Analysis and visualization of professionals linkedin data," in *Emerging Research in Computing, Information, Communication and Applications*. Springer, 2016, pp. 1–9.

[14] N. Kaur, "A combinatorial tweet clustering methodology utilizing inter and intra cosine similarity," Ph.D. dissertation, Faculty of Graduate Studies and Research, University of Regina, 2015.

[15] J. MacQueen et al., "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.

[16] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[17] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review* E, vol. 69, no. 2, p. 026113, 2004.

[18] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical review* E, vol. 69, no. 6, p. 066133, 2004.

[19] B. Everitt, S. Landau, M. Leese, and D. Stahl, "Cluster analysis (> wiley series in probability and statistics)," 2011.

[20] J. Scott, *Social network analysis*. Sage, 2012.

[21] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical review* E, vol. 76, no. 3, p. 036106, 2007.

[22] G. Karypis and V. Kumar, "Multilevel graph partitioning schemes," in *ICPP (3)*, 1995, pp. 113–122.

[23] Z. Lu, X. Sun, Y. Wen, G. Cao, and T. La Porta, "Algorithms and applications for community detection in weighted networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 11, pp. 2916–2926, 2015.

[24] R. K. Behera and S. K. Rath, "An efficient modularity based algorithm for community detection in social network," in *Internet of Things and Applications (IOTA), International Conference on*. IEEE, 2016, pp. 162–167.

[25] I. Der´enyi, G. Palla, and T. Vicsek, “Clique percolation in random networks,” *Physical review letters*, vol. 94, no. 16, p. 160202, 2005.

[26] D. Chen, M. Shang, Z. Lv, and Y. Fu, “Detecting overlapping communities of weighted networks via a local algorithm,” *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 19, pp. 4177–4187, 2010.

[27] A. Clauset, M. E. Newman, and C. Moore, “Finding community structure in very large networks,” *Physical review* E, vol. 70, no. 6, p. 066111, 2004.

[28] B. Bollob´as, *Modern graph theory*. Springer Science & Business Media, 2013, vol. 184.

[29] P. Zhang, J. Wang, X. Li, M. Li, Z. Di, and Y. Fan, “Clustering coefficient and community structure of bipartite networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 27, pp. 6869–6875, 2008.

[30] P. Pons and M. Latapy, “Computing communities in large networks using random walks,” in *International Symposium on Computer and Information Sciences*. Springer, 2005, pp. 284–293.

[31] F.-Y. Wu, “The potts model,” *Reviews of modern physics*, vol. 54, no. 1, p. 235, 1982.

[32] M. E. Newman, “Finding community structure in networks using the eigenvectors of matrices,” *Physical review E*, vol. 74, no. 3, p. 036104, 2006.

[33] W. W. Zachary, “An information flow model for conflict and fission in small groups,” *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, 1977.

[34] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, “The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations,” *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396–405, 2003.

[35] S. R. Chintalapudi and M. K. Prasad, “A survey on community detection algorithms in large scale real world networks,” in *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on*. IEEE, 2015, pp. 1323–1327.

[36] M. A. Russell, *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. ” O’Reilly Media, Inc.”, 2013.

[37] *M. Holub, Twitter Data Mining, 1st ed. Copenhagen: Aalborg University, 2016, pp. 1-95.*

[38] "*The Power of LinkedIn's 500 Million Member Community*". *blog.linkedin.com.*

[39] *Hempel, Jessi (July 1, 2013). "LinkedIn: How It's Changing Business". Fortune. pp. 69–74.*

[40] C. C. Yang and T. D. Ng, “Analyzing content development and visualizing social interactions in web forum,” in *Intelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on*. IEEE, 2008, pp. 25–30.

[41] L. Bijuraj, “Clustering and its applications,” in *Proceedings of National Conference on New Horizons in IT-NCNHIT*, 2013, pp. 169-172.

[42] M. Kaya, O¨ . Erdogan, and J. Rokne, From Social Data Mining and Analysis to Prediction and Community Detection. Springer, 2017.

[43] J. Jackson, "LinkedIn restricts API usage", *PCWorld*, 2017. [Online]. Available: http://www.pcworld.com/article/2883992/linkedinrestricts-api-usage.html.

[44] "NLTK Book", *Nltk.org*, 2017. [Online]. Available: http://www.nltk.org/book/.

# LIST OF PUBLICATIONS

[1] N.Garg and R.Rani "A Comparative Analysis of Community Detection Algorithms using R", in Proceedings of IEEE International Conference on Computing Communication and Automation (ICCCA2017), Galgotia University, Noida, May 5-6,2017. **[Accepted]**

[2] N.Garg and R.Rani "Analysis and Visualization of LinkedIn data using Hierarchical Clustering", in Proceedings of 2nd International Conference on Recent Innovations in Computer Science and Information Technology (RICSIT-2017), UIIT, Shimla, May 19, 2017. **[Accepted]**

[3] N.Garg and R.Rani "Analysis and Visualization of Twitter Data using k-means Clustering", in Proceedings of IEEE International Conference on Intelligent Computing and Control System (ICICCS 2017), VCE, Madurai, June 15-16, 2017. **[Accepted]**