Design and Develop a Framework for Social Network Analysis

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

Master of Engineering

in

Information Security

Submitted by

Navpreet Kaur

Roll no 801433018

Under the supervision of

Dr. V.P Singh
Assistant Professor

Dr. Maninder Singh
Associate Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA 147004

July 2016

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, *"Design and Develop a Framework for Social Network Analysis"*, in partial fulfillment of the requirements for the award of degree of Master of Engineering in Information Security 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. Maninder Singh* and *Dr. V.P Singh* and refers other researchers 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.

Navpreet

(Navpreet Kaur)

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

(Dr.V.P Singh)
(Assistant Professor,CSED)

(Dr. Maninder Singh)
(Associate Professor, CSED)

Countersigned by:

(Dr. Maninder Singh)
Head,Computer Science and Engineering Department
Thapar University, Patiala

(Dr. S. S. Bhatia)
Dean (Academic Affairs)
Thapar University, Patiala

ACKNOWLEDGEMENT

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life. It is a great privilege to express my gratitude and admiration towards my respected supervisors Dr. Maninder Singh,(Head and Associate Professor of Computer Science & Engineering Department) and Dr.V.P Singh(Assistant Professor of Computer Science & Engineering Department). They have been esteemed guides and great support behind achieving this task. I also thank my supervisors for their time, patience, discussions and valuable comments. Their enthusiasm and optimism made this experience both rewarding and enjoyable. I am also heartily thankful to Dr. Jhilik Bhattacharya, PG coordinator, for motivation and providing uncanny guidance and support throughout the preparation of the thesis report.

I will be failing in my duty if I do not express my gratitude to Dr. S. S. Bhatia, Dean of Academic Affairs, for making provisions of infrastructure such as library facilities, computer labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in the field.

I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, which made my stay at Thapar University memorable. Last but not least, I would like to thank my family for their wonderful love and encouragement, without their blessings none of this would have been possible.

Navpreet Kaur

(801433018)

ABSTRACT

Now a day everything around the globe is connected via networks like information, places and events which make a tangle of connections. The outgrowth and favoritism of online social network made available a large amount of data all of a sudden from social organization, human behavior and their interaction. Analyzing social network is to make sense of these complex connections. This work represents the framework to analyze twitter social media tweets using NetworkX and Twitter API. Python language tool IPython/Jupyter is used to examine the networks by applying visual analytic techniques like degree centrality and betweenness centrality to the dataset of twitter hash tags which provides an easier way to analyze the network connections. This framework describes methodology to diagnose each tweet for identification of certain pattern as 'who talk to whom about what'and 'most influential person'in the interconnected/attached network.

Keywords- Social Networking Analysis, Centralities, Twitter, Python Framework

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Network

Whole world is in the form of network. It is a set of nodes or vertices connected by edges or links. Types of network:

1.1.1 Social and Economic Network

It is a group of individuals or set of individuals containing some type of contact patterns between them of interaction. Some of these types of network are: Business Relation between various organizations, Facebook, Twitter.

1.1.2 Information Network

It is type of network that are connected via information. Some of them are: World Wide Web (that are network of webpages that contain the link from one webpage to another webpage in the network), citation of academic papers in the network.

1.1.3 Technological Network

It is type of network that are typically designed for service or commodity distribution. Some of them are: Temporary networks (that consist of sensor networks, adhoc communication network, autonomous vehicle network), Infrastructure network (that

consist of power grid, internet connection of routers and domains, Transportation network.

1.1.4 Biological Network

It is type of network which consist of number of biological system such as network of metabolic pathways, protein interaction network, food web.

The analytical approach to study the network is to predict the network behavior by studying the statistical properties. To understand the statistical behavior there is need to understand the Graph.

Graph is the representation of network. A graph (N,g) consist of N set of nodes and g set of edges where N = 1,2, ... n and g = $[g_{ij}]$ where i,j ϵ N an n $\times$ n adjacency matrix. Here, $g_{ij} \epsilon \{ 0, 1 \}$ represents that the edge is present between the node i to node j i.e. they both are related to eachother.

If the weight of the edge i.e. $g_{ij} > 0$, then that graph represents the interaction intensity of individual in the graph and that graph (N, g) then known as weighted graph.

If graph edges $g_{ij} \neq g_{ji}$, then that graph is called the Directed graph and if graph edges $g_{ij} = g_{ji}$, then that graph is called the Undirected graph for all i,j ϵ N.

1.2 Social Network

Social network specify the art of interaction between peoples in the form of network consisting of nodes behaving as a person and edges behaving as relationship between those nodes or persons. Facebook, Twitter, LinkedIn are the several types and most visited social network which contain a large data about the users and their relationships by going popularity in a short span of time.

1.3 Social Network Analysis

Social network analysis is the technique of research that mainly focuses on comparing and identifying the relationship between and within the group, individuals and system at the eve of real world environment. Social network and their analysis had been extracted from graph theory. Since, graph topology is same social network therefore analysis of graph and social network both are same. Basically, it is defined as study of relationships among human by means of graph theory. It is mapping of flow and relationship between individuals, groups, organization and other entities of information/knowledge. After the visibility of social relationship and knowledge flows, these can be compared and measured. These results can hold the benefit to department, individual and organization. They identify the central roles of person in the network (knowledge brokers, information manager, thoughts of leader, etc.). It aims at exploring the informal relationship such as who shares with whom and who knows whom. This allows visualizing leaders and detecting the diverse relationship.

1.4 Twitter Social Network

Twitter is the most appropriate social network which spread the user information in much fastest way. Two major properties of twitter that make it first choice of data collection comparing to other social networks such as Facebook, LinkedIn etc. are:
First, the spread of information methods of twitter are patent. Twitter users can send and receive short message called tweet of 140 character length. The messages are mostly embedded of plain text and without complex content such as videos, music in it. Therefore, make user to understand the short message in much easier way and to depict whether to retweet the message or not. This core function of twitter social media show that twitter work on simple social awareness stream model[28].

Second, twitter is having more spanned social network structure[28] as it allows it users to follow any other user called friend without validation and approval of the

friend to be followed. So therefore any message posted by user is public to the user followers unless private mode in privacy is not restricted. Therefore this property provides the information to spread as vast as possible.

1.4.1 Methods of information diffusion on twitter are:

1.4.1.1 User Based Information Diffusion

User can share information on twitter using two methods tweets and retweet. The details are:

- **Tweet** user can use this way to originate their own information and share it with their followers.Figure 1.1 shows the Tweet Example. User can post this information by two ways:
  - **External** In this twitter users who are viewing something outside to twitter on some other website and they want that to share with their followers, they can directly tweet from that external webpage and call the JavaScript API of their twitter account. Most of the website support this approach and become popular way to spread the information which user like with their followers.

Figure 1.1: Tweet Example

– **Internal** In this type, user login to their account of twitter and can share their information by inputting the information in the status field and click tweet to post the information.

- **Retweet** this is mainly sharing of that information with your followers which is already tweeted by someone by adding some comments or not with it. This is also the most popular way of scattering information which can be done simply by clicking retweet button.Figure 1.2 shows the Retweet Example

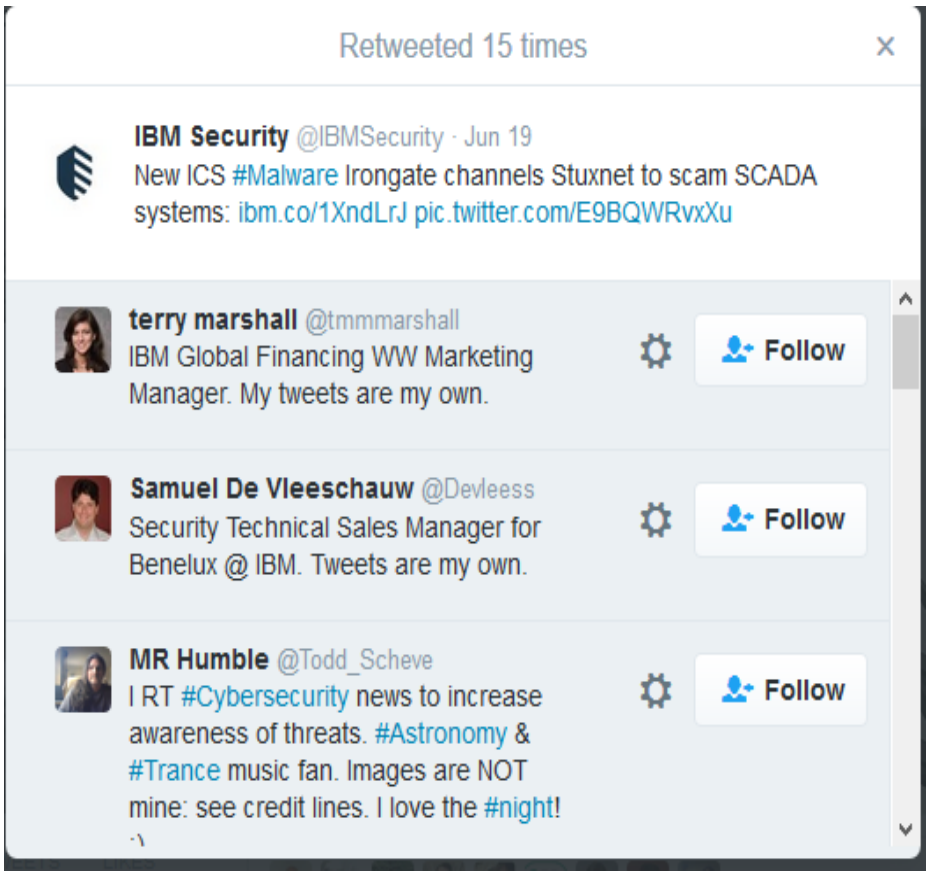

Figure 1.2: Retweet Example

1.4.1.2 Topic Based Information Diffusion

This is a way in twitter where user can spread as well obtain information by using a hashtag append with their topic name. If a user wants to talk or access the tweets

on a particular topic, use of hashtag there comes into account. For accessing the #tag topic related more information. User need to click the tag, whole information about that topic by any other users can be seen. It removes all barriers of scattering information that somehow cant be propagating among strangers. In this unrelated to the relationship of users, the information transfer randomly to all. The Figure 1.3 shows topic based example.

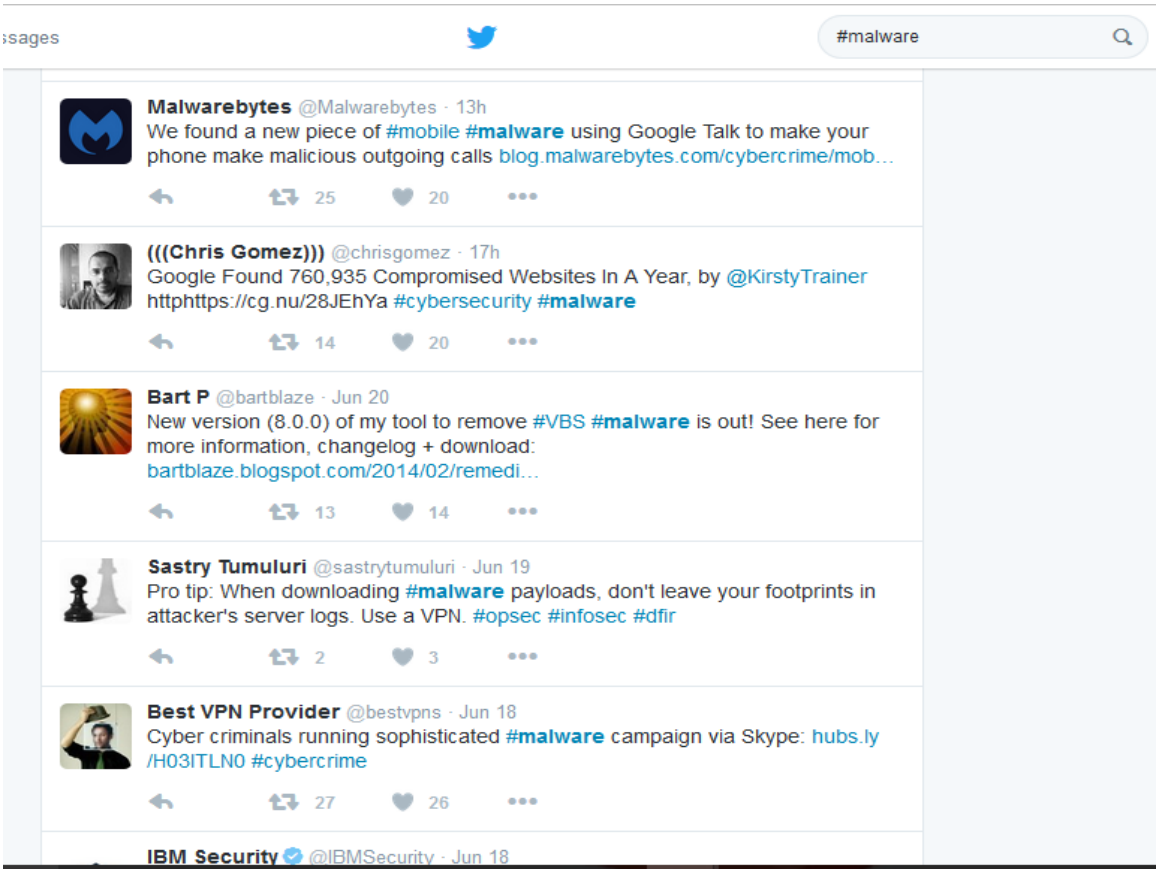

Figure 1.3: Topic Based Example

1.5 Data Collection

For social network analysis data play a quite critical role and for its researcher, collection of data is a big challenge. Usually, data for social network analysis is collected by web crawling.

Web crawling is an automated script or program which surfs the World Wide Web in an automated or a methodical manner. Many authorized sites make use of web crawling to provide quick and up-to-date data to the user. It basically, deal with large data-sets, user can built their own crawlers to crawl the webpages.

Due to new developed data encapsulation techniques and various webpages format data collected by web crawlers may be inaccurate and incomplete. Now, a days a new medium calling related API is available for researchers to get data for social network analysis which is provided by many social network service provider social network service provider like Facebook and twitter and many others had open APIs for contributing more apps to the third party developers . As the benefit of which it make researchers to access more data from social media for their social network analysis.

By gathering information from social network one can figure out important nodes from the network. It contains nodes that are users within a network and ties that is relationship between users such as organization profile; friendship etc. which can be extracted from different values of this network graph by calculating centrality which can Trace communities, envisage (visualizing) whole or part of network, study information Diffusion.

Every day, a microblogging service twitter is making available more than 200 million of facts and ideas. This vast dataset catches the attention by number of researchers all around the world and makes them to find out and discover helpful knowledge from the twitter. This useful knowledge gathered from the twitter network is an important need of scientific industry. The twitter analysis shows an interesting facts and information and then applied on many application domains like social behavior [6], sentiment analysis, and political statements [8] and so on.

These are some facts to find out the relationship between users who commenting on an individual topic and find a network of connected users, these are: 1) Who is

a most prestigious user in a network. 2) Who initiated a conversation and actively involved in the conversation. 3) Which are the most favorited tweets (which retweeted often)? 4) Who are the active users in the network? 5) The timelines on the particular topic in the tweets and 6) The users who shares similar concepts and can be grouped etc. for making a number of important decisions there facts are useful. Facts like advertisements publicity, identifying active users, to find favorite and impactful opinions on an individual topic and find out a driving person in the network [12].

To answer the above raised questions and to analyze the network for the individual topics, the information of tweet metadata is taken along with the text in the tweet. The twitter API is used to download the data. It creates an undirected and directed graph between tweet users and shows important facts between users by using visual techniques. A node is used to represent a user and the edge between two nodes displays the flow of conversation.

To utilize the twitter API for accessing of the data from the twitter the primary framework of social network API need to understand first, which is same for all social network service providers.

1.6 Twitter API

1.6.1 Authentication

API totally depends on the service that they get provided by the website. There is requirement of authentication for API request with the control of information sharing and governing all APIs call by the service provider. The standard approach of authentication use security questions and passwords. Even though, passwords are highly vulnerable to theft. Also, on the other hand there is highly clash and risk between the resource sharing and account security of the people in the social network analysis.

Since, there is occurrence of possibility that during the resource sharing, intentionally or unintentionally someone account privacy may have a chance to leak which is highly unsafe. To make a safe procedure twitter had develop a new protocol name as OAuth i.e. open authentication. This OAuth will help in concealing account security information of user and tokenizing user privileges. So, without providing account information that leads to privacy security checks, OAuth protocol allow users or application which are considered as consumer to access the information. Thus OAuth in twitter reduces the chance of conflict between resource sharing and account securities.

Twitter is providing two authentication models namely, application-user authentication and application-only authentication for different purposes. The main difference between both the models is whether the user context involved in or not. In application-only authentication, on behalf of the application itself, application issues the authorized request whereas in application-user authentication, on behalf of a specific user, application issues the authorized request. These two models saving the data collection time by increasing the API rare limits, thus enable consumers to use different-different identification during a call to the twitter APIs. Thus conclude that application-only authentication act as identification to use an application develop by developer directly and application-user authentication act as identification to use researchers own account of twitter, no matter which identification is used. In regular, 3-way handshaking are used by OAuth to access to twitter API.

Step 1: Application registration

Application need to get register with twitter, if it wants to access the API. To register the application, application owner need to visit dev.twitter.com control panel. Where owner of application also have the privileged to generate an OAuth access token key and access token secret key. If consumer doesnt own an application, then he or she can create new application by login with twitter account to apps.twitter.com and after the application get created. Developer can generate the token and can access the API.

Step2: Issuing consumer API key and consumer secret key

After the registration of the application, two strings are issued by twitter name as consumer API key and consumer secret key. These key can be regenerated and can be revoked again as they are unrelated to user account security thus acting as remedy for the drawback in conventional authentication approach. With these key applications authenticate themselves to twitter social network and can further request tokens.The Figure 1.4 shows API access key of Twitter network.

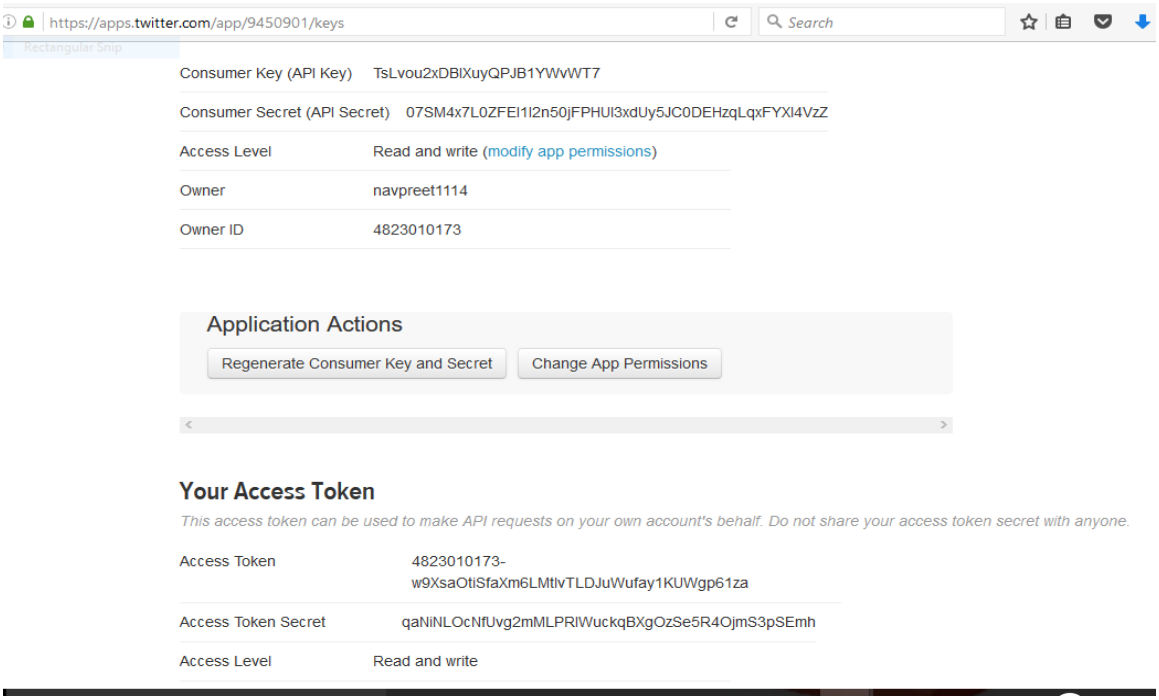

Figure 1.4: API access keys

Step3: Token request

Application can request access token and access secret key by using consumer key and consumer secret key. These both strings can also be regenerated and can be revoked again. These token gives the authorized access to the application for twitter API.

1.6.2 API Usage

Twitter is offering us various types of APIs which covers all the operations on the webpages, which include followers list, updating status, posting new messages etc. the list of all APIs of twitter are present at https: //dev.twitter.com/stream/public. The different types of APIs are REST API, Search API (part of REST API), and Stream API. In this study, Framework is mainly focusing on analyzing social network. So only streaming API was applied and hence introduced below. Streaming API give

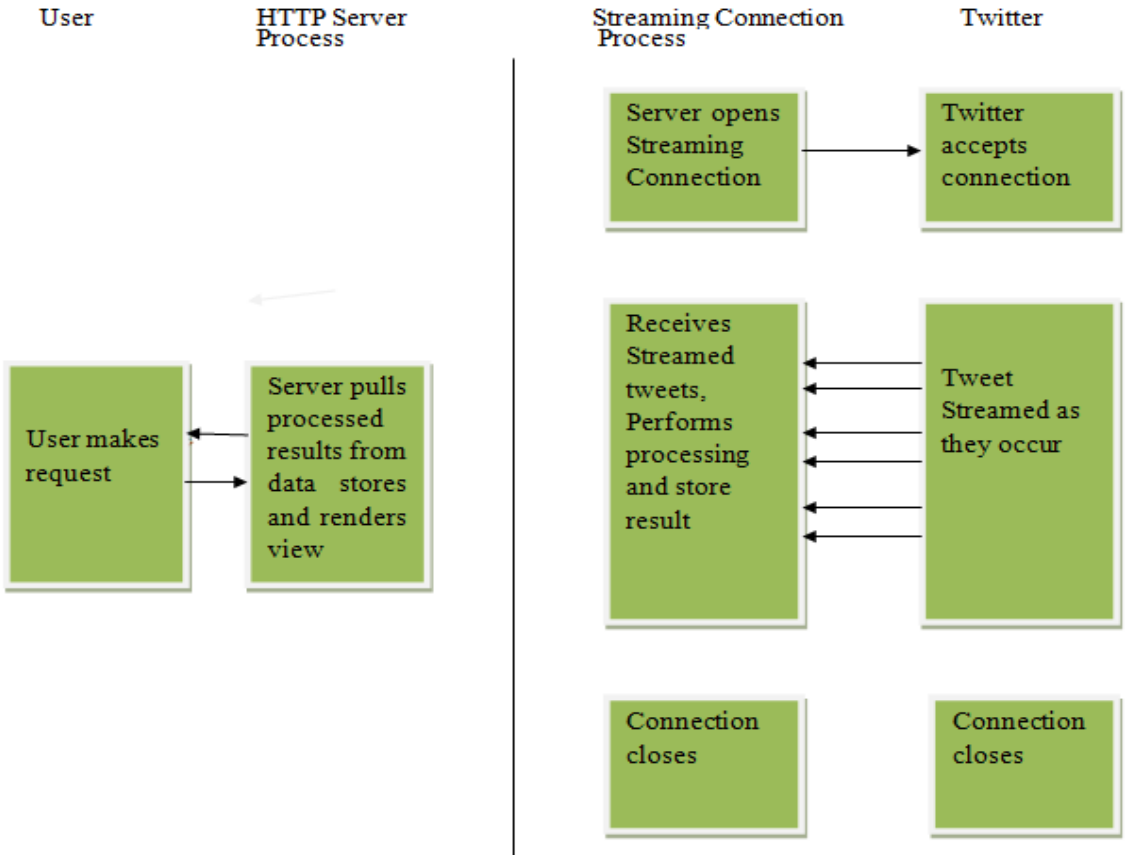

Figure 1.5: Streaming API Flow

access to all tweets as soon as they get published on the twitter. All on an average, around 6000 tweets get posted on twitter within 1 sec and user can get access to only a small proportion to those tweets i.e.$\leq$ 1%. This steaming API is able to give only real time tweets.

If the application is connecting with streaming API, then keeping a persistent HTTP connection open is compulsory. It pushes the data to client regularly whenever it get available as there is no need of client to make the request for data to server again and again for newer data.Fig 1.5 shows the Streaming API flow.

The application which maintain the streaming request run separate from the process which is handling http request because the app which connect with streaming API is not able to make a connection in response to user request.

1.6.2.1 Streaming message type

- Public stream- streams of public data flowing through twitter. It is suitable for following specific topic or users and data mining.

- User stream- stream of single-user data flowing through twitter. It contains the whole data related to the single users view in twitter.

- Site stream- streams of multi-user data flowing through twitter. They are intended for servers which connect to twitter on behalf of many users.

1.6.2.2 Streaming API request parameters

Below are the various types of request parameter which will define the type of data that can be accessed from streaming API endpoint.

- Delimited- Setting delimited to the string length indicates that the status should be delimited in the stream. So, that consumer knows how many bytes to read before the end of the status. Statuses are represented by a length of newline in a byte.

- Filter level- Setting this parameter returns the required attribute from the twitter in the streams getting downloaded. Setting this to default value will return all tweets in streams without filtering specific keywords.

- Language- If this parameter is not set; then the receive tweets may be in any specified language. Making this parameter to language = en will only provide tweets in English language.

- Track- In this parameter, what tweets needed to deliver by the stream is delivered, by the phrase separated by comma in the script and a phrase will match if all the terms are present in the tweets. In this, logical ORs are taken as commas and the spaces are taken as logical ANDs. Phrases are checked if specifically related to text for hashtags, Medias and screen_name, expanded_url and so on. Each phrase must be of 1 to 60 bytes in length.

- Location- In this parameter, a comma-separated list of longitude, latitude pairs specify only geolocated tweets but not the user-location field.

- Count- When reconnecting to streaming endpoint, if count parameter get included, it can help to return backfill missed messages which get occurred during the disconnected period. The value of the count parameter can be set between 1 to 150000 or -1 to -150000. If positive value is provided to the parameter, it will get the live stream values after providing the backfill stream to the client. Whereas if negative value is allotted to the parameter, it will disconnect once by providing only the backfill stream to the client.

- Follow- In this parameter, it will provide only those tweets whose user-id is indicated in the script by comma-separated. For each user, who have specified, the stream will contain

  - Tweet created by user,

  - Tweet which are retweeted by user,

  - Replies to any tweets created by user,

  - Retweets of any tweets created by user.

1.6.2.3 Public streams

Once applications establish a connection with the endpoint of streaming API, they would deliver data that are tweets, without concerning about the Rest API rate limits. Endpoints:

- Post statuses/filter- It return public statuses in stream that exactly match one or more filter predicates. A multiple parameters can be added with it by making user to use single connection with the streaming API. POST request is to be used with it to avoid long URLs. The default access allows user to track 400 keywords, 5000 follow user ids and 25 location boxes. In this the track, locations and follow field need to be considered to combine with OR operator, among them at least one is compulsory to be followed whereas delimited and stall-warnings are optional in it.

- Get statuses/sample- It returns a small random sample of all the public tweets. As the tweet returned by the default access levels are same. So, if two different clients connect to the endpoints, they will see the same stream of tweets. Here delimited and stall-warnings parameter are considered and both are optional in use.

- Get statuses/firehose- It requires a special permission access. Few applications required this access level. It returns all public statuses and creative use of various access levels and other resources can satisfy every use case. Here count, delimited and stall-warnings parameter are considered and all are optional in use.

1.7 Programming Language

The python programming language turn to be so favorite among researchers and data analysts by its well-off collection of data analysis packages (open source) and a wide developer organization. It is a superior tool for cleansing, handling, visualization and

analysis of data. Packages present like NetworkX and graph tool are helpful elements to analyze the workflow of a network. Python is having vivacious and strong-growing ecosystem of packages that helps to fulfil features like drawing and linear algebra by NetworkX.

NetworkX is Python language software package used in programming language for formation, manipulation and to understand the function and design of convoluted network. NetworkX can analyze network structure, can create network models, and can develop different types of random and classic network. It also has language data structure for graph, Digraph and Multigraph. It is library which get use in Python and can handle one million of nodes. It fulfills data structure for networks or graphs including drawing tools and algorithm generators for graph. A function provided by the network API that uses a graph object as an argument. The graph object methods are restricted for handling and reporting. Ulrik Brandes [1] proposed algorithm which is being used by NetworkX for calculating Betweenness Centrality. This algorithm is for large networks that unite with the Traversal algorithms used for solving problem of single source shortest path by reducing both space and time complexity. Brandes uses dependency accumulation to find the betweenness centrality, which access the vertices in the reverse order of the Breadth First Search (BFS) access method. This dependency accumulation makes this algorithm faster among all previous ones. This algorithm comprises of four stages [29] :

Stage 0: Local initialization

In this stage, betweenness centrality score to the each vertex is initialized to zero and stage 1 and stage 2 get computed for each vertex with an iteration method.

Stage 1: Global initialization

In this stage, the data structures that will be used in Stages 2 and 3 are initialized that includes a queue, stack and three arrays. The first array from each vertex counts the number of shortest paths to the root of the current shortest path tree. The second

array from the root measures the distance of each vertex. The third array is an array of linked lists. Initially the distances of all vertices from the root are set to .

Stage 2: BFS traversal

In this stage, BFS is computed from a given root that finds the shortest path to all other vertices. In this each element is placed in a queue when it is found that is later placed in the stack when it is dequeued from the queue. As part of the BFS traversal the distance from the root vertex, to each vertex is computed. For each vertex, found in the BFS traversal there is a list of parental vertices that are all one hop closer to the root. Thus, all those vertices shortest paths go through its parents and these are accumulated in one of the variable array.

Stage 3: Dependency accumulation

It computes betweenness centrality using the dependency accumulation technique of Brandes [1] by calculating the pair dependency between the pair of vertices. This pair dependency follows the recursive relation for all the pair of vertices.

1.8 Node Distribution

The main aim of the social network analysis is to find the importance of the node in the network or analyze the leverage of a node over the network. The main perspective to analyze the leverage of a node in social network analysis is the centrality study, whose main objective is to analyze whether a node in a network is at advantage or disadvantage position in the network structure.

1.8.1 Degree Centrality

The node with more connections to other nodes is in advantageous position for catching whatever the flow is in the network. This flow can be information, rumors or might be viruses also. This independent nature of the node on other nodes makes them more powerful in the network.

Degree Centrality is based on the fact that which node is having the most connections with other nodes in the network. To determine those connections, it adds up the edges weight with the rest of the nodes. In directed network (where edges have direction), commonly degree centrality measures separately, known as Indegree and Outdegree.

In directed network, Indegree refers to number of connection made to the node from whole network. Outdegree refers to number of connections made from the node to the whole network. Edges are related to some positive sides, like collaboration or friendship, Outdegree is frequently interpreted by importance and Indegree by its popularity. It is used to find so connected users, favorite users who used to hold most of the information or users who quickly make connections with wider network. It is easiest way to measure node connectivity. Most of the time its helpful to look at Indegree (no of inbound links) and Outdegree (no of outbound links) as different measures, an example is by lookout data transaction or account statement activity.

High Outdegree nodes are the individuals who are able to exchange their information within the network and also to explore their point of views to others. Mainly Outdegree deal with leverage and Indegree deals with the popularity.

Centrality attribute of an individual can be measure by counting in and out connections of the individual which shows who is at more central position in the network. The centrality attribute can be measure of the whole graph which will measure how centralized the graph is by raising the value of centrality with group of actors in the network. The general definition of centralization was proposed by Linton Freeman in 1979 [26]. The originate idea of it was that the individual with more value of degree centrality are more leverage.

Here, Bonaich proposed the modification of degree centrality in [27]. Bonaich argue that the two nodes having high degree doesnt state that these are important nodes in

the network. Bonaich going step further and says that nodes that are well connected in the network that connects to other doesn't make it powerful but of course central as information may reach to them by many others ways. Whereas the nodes that are not well connected in the networks but connected to others makes them powerful, as theses nodes will depend on you, Bonaich proposed that both power and centrality were the functions of connections of one neighborhood.

1.8.2 Closeness Centrality

Degree centrality may get deprecate because it takes only direct connection of the individual or the connection from the node neighbors, else than indirect connection to all others. That means one individual may be having large number of connections to others but others might be all disconnected from the network which makes that individual quite central among neighborhood.

Here, closeness centrality highlights the approach of calculating distance of the individual node from all the other nodes in the network. In graph theory, distance between the pair of nodes is calculated by length of their shortest path [26].

Closeness is the inverse of farness, where farness is defined as the sum of distance from all other nodes which conclude that lower the distance from all others nodes, more the node is central.

When an individual node is more reachable by other individual or able to reach other individual at shortest path length, that position is called favored position. That favored position act as reference point as these nodes are closer to more nodes than all other in the network. Closeness gets measured by how fast an exchange of information take place from reference point to all other individual nodes in the network.

1.8.3 Betweenness Centrality

Closeness centrality provides how fast nodes can reach to other in the network by which it can easily access the data flowing in the network. But it doesnt indicate how much control a node has on the information that is flowing in the network i.e. how frequent that node is coming between the path of other nodes. Freeman introduced betweenness centrality to quantify the role of individual on the information between the other individual in the social network. Some individual plays a contemplate role between other individual communication which gives benefits to them in term of power.

Betweenness centrality is described as ratio of smallest path through a node or edge to the total number of smallest path in a network which give us a central node among all other nodes in the network. This measures how many times a nodes fall on the shortest path between rests of the nodes. It is used to find out the users who make a flow impact around a system. It used to analyze the dynamics of communication. A high betweenness add up could point someone hold on the authority above or control the collaboration among different clusters in a network or on the boundaries of both clusters.

1.8.4 EigenVector Centrality

Freeman introduces the third centrality known as Eigen vector centrality, it aims to find the top central node in the network within global network which focus less on local distribution of node pattern [26]. It defined Eigen values on the basis of factor analysis. It defines dimension [25] of distance among nodes. The location of each node with respect to its dimensions is called Eigen values and grouping of these values called Eigen vector. The first dimension capture the global aspect of Eigen vectors then second and then further local aspect of values get capture.

This centrality act as a recursive [25] version of degree centrality as, the node centrality is proportional to sum of centrality to which it has connected. This centrality measure the popularity of the node within the network. It calculates the importance of nodes in the undirected graph whereas page rank algorithm make use of Eigen vector centrality but measure the importance of node in the directed graph.

Chapter 2

Literature Review

This chapter provides systematic review in the field of Social Network Analysis exploring various Network Centrality techniques.

2.1 Systematic Review

Analysis of Social Network is very important topic over the World Wide Web. The research in this area is unstoppable. This survey has been stated in the area of Social Network Analysis from year 2006 to 2015.

Danyel Fisher et. al [2] presented a method that applied on network analysis for supporting the task of different authors in Usenet group. They visualized and computed networks created by replying patterns for each author in selected newsgroup and found that second-degree networks gave them clear differences among different types of authors and newsgroup. Their result showed that newsgroup changes by means of population of contestants and the roles that they were played. This approach had applications for both researchers finding to characterize many types of social cyberspaces and contestants finding to differentiate interaction between content authors and partners.

Marc A. Smith et al. [3] presented a toolkit named NodeXL for network overview.

Discovery and exploration to implement as add-in to the Microsoft excel 2007 spreadsheet software. They also presented NodeXL data analysis and visualization features with the data sample of social media got from an enterprise social network. A NodeXL operation from import the data to network statistics computation and refinement of network visualization through sorting; clustering and filtering is defined here. Sociologically relevant differences in interconnection patterns between different employee participants of social media reveal under these operations.

Danah Boyd et al. [4] examined the way of retweeting practice by which individuals could be in a conversation. Individual participants retweet for diverse reasons using different-different styles, as retweeting has become a convention inside the twitter. They pointed how different communicate quality authorship are mediated in different ways. By using many types of different case studies and observational data, they find out that retweeting as a conversation practice.

Haewoon Kwak [5] studied the topological characteristics of twitter and its information sharing power as a new medium. They analyzed the tweets of the trending topic and user participation on their behavior. They classified the topic on the basis of active time period and so the tweets by showing the majority of topics are live news or persistent news in nature. By looking at the retweets, they discovered that any retweeted tweet reached around 1000 users by not considering the number of followers of the original tweets.

Meeyoung Cha et al. [6] measured the impact of users by using three measures; these are total number of followers, number of time user name shown on the text and number of retweets. They reveal the fact that among all, celebrities are the most followed users. They mainly focused on the Directed links in the twitter are network that determine the influence of users on other by the flow of information. Collecting the large amount of data from twitter, it mainly focused on the three major concepts that are Indegree, retweet, and mentions. They give a detailed study of users influence

based on time and topic. It results in several outputs like most influence user holds the influence over variety of topics. Users influence can be gained through tweeting regular to a particular topic. People having high degree by mentions or retweets are necessarily not influential.

David Ediger et al. [7] presented a graph characteristics toolkit named as GraphCT for representing a graph of massive social network data. Cray XMT, GraphCT evaluates the degree distribution and betweenness centrality of an artificially generated (R-MAT), millions of vertex, billions of edges graph in 55 minutes. They used GraphCT to analyze public data from twitter by taking in account of two main case studies held in 2009. The gather the tweets of uneza H1N1 in September 2009, Atlanta flood tweets in September 2009 and also all public tweets of 1st September 2009. The manage connections of twitter appears in a tree-structured as a dissemination system. Using GraphCT, they ranked actors within conversations and help analysts to focus an attention to the small data subset.

Michael D. Conover et al. [8] described some methods to predict to alignment of politics of users (twitter users) on the basis of political communication content and structure during 2010 U.S elections. By applying latent semantic analysis on users tweets context, they find out the unveil structure of data linked with political affiliation, but unable to find the topic detection that may improve predict performance.

Marcin Mincer et al. [9] addressed an issue associated with a social network analysis application for analysis and investigating social relationships of people. It deals with the certain types of outcomes from the social network. That is to determine the interpersonal connections on the social network, to determine the importance of actors in a given social network and to detect the communities of the people. The methods and models objectives were to investigate the applications of social network analysis techniques like data mining for two social networks. Data is collected from Facebook by using its user profile and Twitter by using two most used #tag i.e. #Justin Bieber

and #Oslo massacre. They provided a short introduction for representation of social network. The analysis on the social network is performed using centrality measure of betweenness, closeness and eigenvector values. They also discussed the strength and weakness of social network analysis techniques.

Lea Ellwardt et. al [10] argued by taking social network perspective that social status and group boundaries in the casual work area network concluded who the objects of positive and negative gossip were. These networks were collected between 36 employees in a public child care organization, and determined using exponential random graph modeling (ERGM). As imagined, both positive and negative gossip focuses on contemporaries from the own work group among these, negative gossip is fairly targeted, with specific individual objects. On the other hand, positive gossip was spread more all over the network.

Naveen Sharma et. al [11] designed and evaluated who-is-who novel service for concluding attributes under which each Twitter user characterize individually. Their methodology avails the featuring lists that helps a user to group other users who are keen to tweet on a topic that is of interest to them, and collectively follow their tweets. Their key view was that the Meta data lists having name and descriptions provided worthy semantic cues about the users who were included in the lists, included their expertise topics and how they are publicly perceived. So, they inferred a users expertise by getting the meta-data of crowd sourced List having the users. They showed that their methodology could comprehensively and accurately inferred millions of attributes of Twitter users, with a huge majority of influential users of Twitter. Their work provided a foundation to build better search and recommendation utility of Twitter.

Rizwan Mehmood et al. [12] attempted to give knowledge based process discovery on the dataset of twitter containing hashtags and the techniques of visual analytics, whose aim is to give people information in such a way so that people are able to

get the knowledge in the data easily. They have collected the hashtag data tweets using streaming API in the KDD process. The processing of those data tweets was done by XLRD, XLU python modules that manipulate data into excel file which will help in visualization of the processed data in Gephi analysis tool. They also analyzed the text in the tweet and that of metadata linked with each tweet to identify the important patterns like "who converse with whom"and "how much"by using Degree centrality and Betweenness centrality values. This research work put an impact of visualization and cluster techniques to find out similar user groups. They also study the measures of social network that reveal the influence of users with respect to the specific hashtags.

Seth C. Lewis et al. [13] discussed an approach for blending manual and computational methods of all the content analysis process that might give more fruitful results. They also drew a case study of twitters news source to elaborate the above discussed hybrid approach. By taking the Andy Carvins Arab Spring coverage as case study, assuming twitter as virtual newsroom of journalistic sourcing. Python script was written to collect the case study related data and also for parsing the data according to date, body text, mention username, RT info from the tweet, that will convert the raw data into CSV file of Excel. The combinations of both manual and computational techniques preserve the strengths of content analysis done by using PHP scripting language. They made the twitter a convenient tool for journalist to interact with possible sources and gather information of large source without even leaving the office. Thus it identifies the two key variables: firstly, to find the type of interaction based upon whether the interaction was retweeting or mention and secondly, the type of source is being interacted with.

Itai Himelboim [14] took a social networks approach to examining the Twitter talk suggested by politically concerned television hosts. Two features of user relations are examined: the users interconnectedness as a sign for users exchange of networks for opinion and information, and disclosure to a political range of information source.

Users prefer exposing themselves suggested by findings for political minded information sources. Moreover, the hosts of television failed to suggest an opinions and ideas exchange between their followers. In compare, other sources of information that stated about these hosts suggested deeper interaction between their followers.

Todd E. Malinicka et. al [15] examined the correlation among structural location (namely, degree centrality) and media coverage. Their central assumption was that the network centrality of social group actors was certainly linked with the popularity of actors cited in the news media. They used two modes of data from a communication network in British Columbia, and examines the relation among their frequency and structural location by that they were cited in news media with respect to particular frames(about environmental protest, forest conservation and related problems). They asked a group of social movement participants about their ties with high profile actors. They also compared the network centrality effects of the target actors and few attributes of the target actors by means of media coverage that received by each of the actors. They found that network centrality is related with media that controls actor attributes.

Nadeem Akhtar et al. [16] mainly focused on the representation of Facebook network to uncover the hidden relationship. By identifying the behavior of high degree in the Facebook network, it gave the conclusion of having greater number of friends in the network. By uncovering own user profile of the Facebook network, it collect the data by data scraping from the Facebook in the .mhrw format. The subgraph are obtained through file handling technique which then get converted in CSV file to open in Gephi analysis tool to visualize its high degree nodes through its various measures of Degree, Eccentricity and Closeness centrality.

Naheed Akhtar [17] presented a comparative analysis by using four different social network analysis tools: that are Gephi, NetworkX, Pajek, and IGraph. It gives us a

detailed study based on graph types, input file format, platform graph features, execution time, and algorithm complexity among all the four tools. A brief introduction on the NetworkX is based on the platform NetworkX is a type of library which runs on python platform with fast computational time and can handle one million number of nodes. It takes O(n.|E|) time complexity. NetworkX library is more useful for the tasks involving millions of nodes and for the difference and the union set of operations between the nodes.

Jari jussila et al. [18] focuses on the use of twitter. The data was captured from the case study of conference CMAD (Community Manager Appreciation Day) held in 2013. The data was taken from the twitter of the conference in three manners i.e. before, during and after the conference take place. In this case study data was collected by the python script by using REST API. The query to the tweets was made by the NoSQL Mango Database. Processing of tweets was done by python script which also converts the raw data into tables of Xml format. The data-driven visual network analysis done by Gephi tool, by which conference participants and conference discussed topic has been analyzed in a network using Betweenness centrality and Degree centrality. They able to find out highly impact topic and discussions, powerful participants of the conference. Therefore, the need of conference information was taken from the information visualization, having interferences to improve the co-organizing of conferences and its planning, also for the use of twitter in conference communication.

Itai Himelboim et. al [19] mapped the Twitter networks of 10 political controversial topics, detected clusters and subgroups of highly connected users and coded links and messages in them for orienting politics. They found that Twitter users were differently to be exposed for cross-ideological content from the user clusters they followed, as these were mostly homogeneous politically. Last, they found that more precise topics of controversy had both liberal and conservative clusters, in case of broader topics, superior clusters reflected sentiment conservatively.

David Burth Kurka [20] described how to approach the computational researches by analyzing such systems by different method. They proposed taxonomy to define and classify different research categories. Each category of research is defined in which the main work, different perspectives and discoveries are highlighted. The mainly focused area of this research was online social network analysis that was combined of three major components such as Structure analysis, Social data analysis, and Social interaction analysis. The author had described each component very well by taking in account the working and use of the component. Social data analysis till now had reach the various research platform of sentiment analysis, prediction of upcoming events, trending topic detections and social recommendation system.

Elizabeth I.Peele [21] focused on ISIS′s twitter propaganda campaign and to discover underlying network structure using social media analysis and twitter own APIs. The resulting scale free network structure is analyzed to see how it affects the twitter ISIS′s dissemination propaganda. Basically, this research concludes hopes that how can terrorist organization can be stopped to spread their message online via network structure.

Berrin Erdogan et. al [22] examined the relationship among member-leader exchange (MLX) quality and monition network centrality using multisource data by a sample of 250 employees of retail and their corresponding managers in Turkey to check their hypothesized model of costs and values of being sought out for advice. Giving on the basis of tenants of generation of network theory, they predicted that the chances of focal actors to help their own tendency and to others would be behavioral of the relationships among MLX quality and their network centrality. Their results revealed that MLX quality is related to centrality only for those who had high chances to help other workers and a low chances to gossip about other workers.

Nia M. Dowell et. al [23] adopted a fresh approach, using context as a tool to look into

its association with two established measures to learn: social centrality and academic performance. They demonstrated how language characteristics diagnostically gave the performance and social position of start learners as they interacted in a MOOC. They used CohMetrix, a theoretically, linguistic modeling tool, to explain students forum across five potent fresh dimensions. Using SNA methodology, they determined learners social centrality. The results indicated that learners performed substantially when they engaged in more style discourse, with deep level adhesive integration and simple synthetic structures. Implications for more research and practice were talked about regarding misalignment among those two learning related outcomes.

Katherine C. Chretien et. al [24] aimed to describe why and how a medical student use Twitter for professional development. They observed tweets of almost 30 students for the duration of 8 months and noted the structured field notes. By sampling, each key informant interviews were conducted to explain Twitter use and values until thematic was reached. Ego network and subnetwork analysis was performed of student key informants. Twitter represented as a professional tool that fulfilled the traditional medical school experience. Purposely, super users approached their use of Twitter and were aware of online professionalism with respect to good Twitter citizens. Two key domains where Twitter provided values were: access and voice. Student got access to information, to a variety of perspectives with patient and public perspectives, and the support by the communities.

Chapter 3

Problem Statement and Objectives

3.1 Problem Statement

Different researchers of different domains have focused on the knowledge discovery using twitter. A prime focus is on the sentiment analysis, political findings, clustering of messages and some real world development. very little research work was done on finding a relation among user sharing opinions for general topics. Also the information that gets discovered was also not visualized properly to get deeper information from the large scale of data.

A lot of data is available on the social networking sites, which depicts the sentiments of the people. However it is not manually possible to analyze this data because of volume, veracity and velocity.

Various tools like Gephi, Nodexl allow us to graphical visualize this data. However, automated solution for social network analysis are still very few and with limited functionality.

3.2 Objectives

The following are prime objective of this work:

- To capture Social Network Data in raw format.
- To Analyze captured data emphasizing various trends.
- To Find correlation among the nodes and their respective data generation.

Chapter 4

Implementation Details

4.1 Twitter API

It is the key provided by twitter to their developers which allows us to interact with the twitter data consisting of tweets and its several attributes. For making request to the Twitter API server side scripting language like ruby, python and php is needed. In this framework python script is implemented to download the tweets from the twitter which would return the results in json format. Json is a compact, text based format JavaScript Object Notation for computers to exchange data.

```
from twitter import Twitter, OAuth, TwitterHTTPError, TwitterStream

ACCESS_TOKEN = '711408546018033669-Lzd2gUMUOFwGkGk1rsaDyN4RIxC6JJk'
ACCESS_SECRET = 'R2bTfZRyReAlZhmt8GQc0913oD6F103NTxXI1L3ko4W3X'
CONSUMER_KEY = 'JfIJN90WOYOI9b9RXpOsmT6PT'
CONSUMER_SECRET = '4M0IRoYQM5Fz3KKkq6DPdVPMIdzisRa2vqhGpQuRLJKDpUcviD'

oauth = OAuth(ACCESS_TOKEN, ACCESS_SECRET, CONSUMER_KEY, CONSUMER_SECRET)
twitter_stream = TwitterStream(auth=oauth)
```

Figure 4.1: Twitter API access

4.2 Gathered Tweets

In social network analysis it allows us to analyze the members in the social network and how they are known to each other. To find relative information about who communicating with whom and about what topic they talk, data from twitter need to be gathered. To access the tweets, twitter Streaming API is used. In Figure 4.2, code of python shown that is used to filter the tweets on the twitter related to #tag malware and that which contain only English language.

```
iterator = twitter_stream.statuses.filter(track="#malware", language="en")
```

Figure 4.2: Filtering Tweets in Real Time

Malware named for malicious software is a kind of software that stops the computer operation, collects sensitive information; take an access to private system or shows non-required advertisements. The study of malware analysis is needful for current crime ware analysis in the enterprise. There are plenty of crime ware variants with so many tricks to conceal their real intent.

Hashtag # is random chosen term which gets invented by users to chase all conversations link to topic event or brand. It eases the search by classifying the tweets. User can put a hashtag in any tweet. Figure 4.3 shows the example of the tweets downloaded by using #tag malware through python script by using Streaming API. It shows that each tweet contain meta-data like screen_name, user_id, tweet text, retweet user_id and so on.

4.3 Target Dataset

It includes focusing on selecting a dataset on which analysis has to be performed. The data is in json format consisting of many attributes out of which many are not of concern. Having much of tweets in a json format where each tweet contains a ton of information but since interested in only who talk with whom about what. This

```
    "coordinates": null,
    "source": "<a href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>",
    "favorited": false,
    "in_reply_to_user_id": null,
    "is_quote_status": false,
    "geo": null,
    "truncated": false,
    "favorite_count": 0,
    "text": "Coders that write malware are known as hackers. Those who write malware to commit crimes are known as \u201cblack-hat\u201d hackers #CodingFacts",
    "entities": {
        "symbols": [],
        "user_mentions": [],
```

```
    "id": 735049467347521536,
    "retweet_count": 0,
    "in_reply_to_user_id_str": null,
    "created_at": "Tue May 24 10:07:04 +0000 2016",
    "contributors": null,
    "user": {
        "id_str": "481882059",
        "utc_offset": 7200,
        "default_profile_image": false,
        "favourites_count": 66,
        "following": null,
        "name": "NerveIT | Web&Apps.",
        "follow_request_sent": null,
        "is_translator": false,
        "profile_image_url_https": "https://pbs.twimg.com
```


Figure 4.3: Snippets of Tweets Downloaded

implementation need only is screen_name and text from each tweet. Therefore there is need to process the each tweets and gather the target data that is shown in Table 4.1

Sr.no	Attributes	Description
1.	Screen_name	It tell the name of the from_user who have tweet
2.	Retweeted Screen_name	It tell name of the from_user who have retweeted someone tweet
3.	Text	The main tweet posted by user containing and #tags

Table 4.1: Attribute to process

4.4 Processing Tweets

It defines type of processing applied on raw data to ready it for further analysis. It converts the data into ease format that can be effectively used for the analysis. The representation and quality of data to use under analysis is foremost. The knowledge discovery become difficult if there is irrelevant, unreliable and noisy information also, due to which error take place. Therefore to process the data for analysis is much difficult process.

This research calculate processings by extracting from_user screen_name, to_user screen_name and hashtag topics. From_user screen_name are taken from the users information of the tweets data, is a person who post or send the tweet. To_user screen_name are taken from text part of the tweets data referenced with @ tag by from_user and also, from the screen_name of retweeted users information is people who get mention in the tweet text or receiving the tweets.Fig 4.4 represents the flow chart for implemented work.

Processing of tweets will take place in two parts:

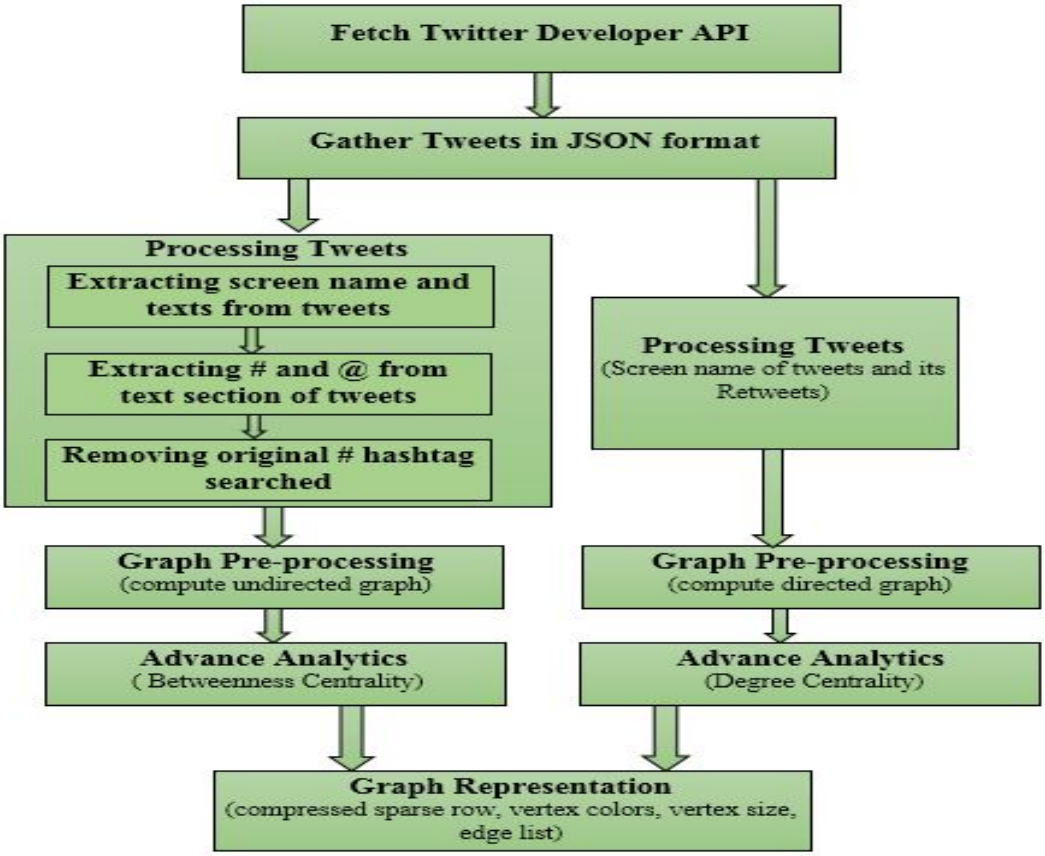

Figure 4.4: Flow Chart For Implemented Work

4.4.1 Processing Tweet Text

In this processing, From_user screen_name are taken from the users information from the data of the tweets i.e. a person who post or send the tweet and To_user screen_name are taken from text part of the tweets data referenced with @ tag as well as with #tag.

Thus, tweet text containing special symbol such as @, #, RT using language of twitter. @ Sign is significant code which is used to refer users on twitter. @ is append with screen_name into tweet to represent a reference to individual or send him tweet. RT refers to retweet in a tweet text. This symbol is code to recognize tweet by other users that its retweet to someone others tweets.

```
['RT', '@actionfrauduk:', 'Microsoft', 'Tech', 'support', 'scammers', 'turn', 'to', 'screen', 'locking',
'malware', 'to', 'dupe', 'victims', 'according', 'to', 'new', 'research', 'https://t.co/Yi…']
['Updated', 'Skimer', 'malware', 'infects', 'ATMs', 'worldwide', 'https://t.co/LGAgcwXoCo', 'via',
'@StackTime']
['RT', '@kafeine:', 'Angler', 'EK', '+=', 'CVE-2016-4117', '(waiting', 'for', 'CVE', 'confirmation..but',
'what', 'else', '?)', 'Post', 'updated', ':', 'https://t.co/0yUzrDqb6p', 'https://t.co…']
['Unmatchable', 'adversive', 'malware,', 'squared', 'off', 'spyware', 'other', 'oppugnant', 'systemic',
'insecticide', 'essentials:', 'ZByadU']
```

Figure 4.5: Processed Tweets containing @, #, RT

Screen_name and text are hold in a list. Screen_name need not to further process whereas processing of text is in need. Therefore, tweets are split into words by space, by which the tweets text is now turns into array of words. These words containing @ and # will be filter from the array. So a multilevel dependency exists between screen_name and text data in a list. These two types of informations from the text make a multivalued dependency with the From_user and get saved in the array of list. Also, to avoid ambiguity from the data set, the original hashtag that was searched during accessing of tweets from the Twitter need to be removed.

```
list2 = []
for item in data["users"]:
    var1 = item['user']['screen_name']
    list1 = [item['text'].split(" ")]
    for subitem in list1:
        subitem = [i for i in subitem if i != '']
        x = 0
        for subitem1 in subitem:
          if subitem1[x] == '@':
            list2.append((var1,subitem1))
```

Figure 4.6: Processing tweet text by '@'

4.4.2 Processing Retweet

In this processing, From_user screen_name and To_user screen_name both are taken from the users information of the tweets data i.e. a person who post or send the tweet. Most probably Retweet information is only present in some of the tweets rather than

```
file = open('malware.json')
pairs = []
data = json.load(file)
for item in data["users"]:
    try:
        pairs.append((item["retweeted_status"]['user']['screen_name']
                      ,item['user']['screen_name']))
    except:
        pairs.append(('no',item["user"]['screen_name']))
```

Figure 4.7: Processing Retweet Information

all the tweets. Therefore, resultant data will consist of only those people screen_name who have retweeted. These two types of informations from the twitter data make a multivalued dependency with the From_user and get saved in the array of list.

4.5 Graph Pre-processing

It is process of converting data from list format to tweet graph format. Tweets graph is the user interaction graphs that are created by adding on edge into the graph for every mention (denoted by prefix @, #) of a user by the tweet author. After finding screen_name and hashtag topics from text, transformation of data into nodes and vertices take place using Python library NetworkX. An undirected graph is generated from the data set of processed tweets by text and a directed graph from the dataset of processed Retweets information.

```
plt.figure(figsize = (20,20))
d = nx.Graph()
d.add_edges_from(list4)
nx.draw(d,with_labels=True,node_size = 100,node_color='yellow')
n = d.number_of_nodes()
e = d.number_of_edges()
```

Figure 4.8: Graph Pre-processing

4.6 Advance Analytics

Centrality is playing an important role that helps to find out the critical positions in the whole network. Central points have been compared with the popularity and leadership opinions. To identify critical positions, several measures of the vertices and its connectivity have been used. Among them Betweenness centrality is calculated for Undirected graph which is created using processed tweets of text and Degree centrality is calculated for Directed graph which is created using processed tweets of Retweet screen_name. In Figure 13, it is clearly shown that a function of centrality is called using python language that will return the values of centrality by using sorted function. This code will take the graph as an input and will return the values of betweenness in sorted manner. NetworkX use the Brandes [1] fast algorithm of betweenness centrality.

```
def centrality_sort(centrality_dict):
    return sorted(centrality_dict.items(),
                          key=operator.itemgetter(1))
between_cent = nx.betweenness_centrality(d)
between_sorted = centrality_sort(between_cent)

print('----------------Betweenness Centrality------------------')

print('highest_degree:' ,between_sorted[-n:])
```

Figure 4.9: Betweenness Centrality Measure

4.7 Graph Representation

To analyze the tweets graph of the twitter network, the color and size of nodes are set according to the values of centrality. So, that greater the value of centrality greater the size of node which give ease in visualizing the results from the graph. Therefore, values of Betweenness centrality are taken along with the node as the parameter by which node_size and node_color get change.

```
plt.figure(figsize = (50,40))
g2=d.subgraph(degree)
size=[]
for i in range(0,len(newsorted)):
    g2.add_node(newsorted[i][0],size=int(newsorted[i][1]*100000))

sz = nx.get_node_attributes(g2,'size')
map_values = g2.degree()
values1 = [map_values.get(node, 0.25) for node in g2]
values = [sz.get(node,0.25)for node in g2.nodes()]

nx.draw(g2,with_labels = True,node_size = values,node_color=values1)
```

Figure 4.10: Centrality Measure

In Figure 4.10 newsorted is a list containing values of betweenness centrality corresponding to each node in it. The values of betweenness centrality for each node is get divided by some tens value, so that node color and size get raised. As large the value of betweenness centrality, large will be the node size. Then a subgraph will be generated by taking in those values, which will give the clear representation for the nodes according to their centrality values in an ascending order.

Chapter 5

Results and Discussion

Visualization of information from the twitter data provides the sight of network that people communicating over the #tag malware. From which, network of people communicating and their topic are derived and hence, back with the result of most discussed topic and most influential people.

A python script was implemented to find the screen_name of the person who tweet, screen_name of the mentioned twitter users, screen_name of the user who retweet and the hashtags used in each tweet. Users of twitter can be presented as nodes of network that are interconnected through retweets and discussion. Two type of network graph is generated.

First network graph is created by using NetworkX of python which include two types of nodes, consisting of twitter users and hashtags i.e. when one has mention other that connect the pair of users, the one who mention and the one who got mention. Also, users are connected to the hashtags that are used to mention and also to the hashtag that they have used to tweet. Second network graph represents the interconnection between the people communication over the twitter via retweets i.e. the person who retweet to the person whose tweet get retweeted.

Since, the Centrality indices are important tool that are described on graph vertices to analyze the social network. They are built to interpret the prominence of users according to their grade by their position in the network. Therefore the python script was implemented to find the betweenness centrality of the undirected network graph to find 'who talk to whom about what'and degree centrality of the directed second network graph to find "most retweeted person'by max Outdegree value and the 'most influence speaker'who make maximum reply to other by max Indegree value.

5.1 Who talks to whom about what?

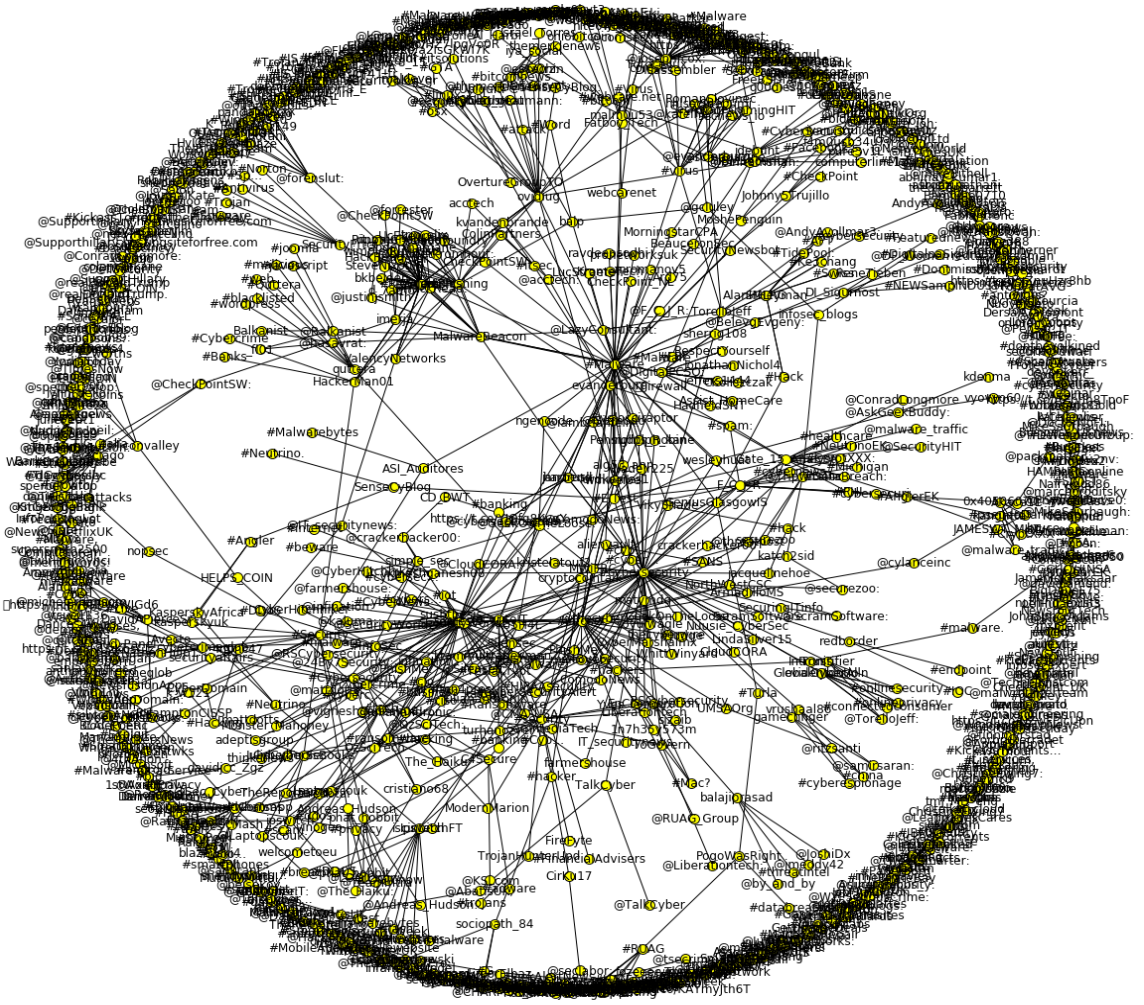

Figure 5.1: Undirected graph between users and text

In undirected graph edges are created between the from_user screen_name and there text containing @ screen_name and # topics. An undirected graph is a graph whose edges are without direction and multiple edges are ignored between two nodes.Figure 5.1 represents the Undirected Graph Between the users and text.

Betweenness centrality gives the shortest path among the nodes and also V number of times a node act as a bridge. Measure how many times a node act as a bridge. It was presented for measuring the control of a person while communicating with other person in a social network by Linton Freeman. In this category, a high probability vertex can occur between a shortest path which is randomly chosen and these randomly chosen vertices have a high betweenness. To compute the betweenness of a vertices V in a graph G = (V, E). Do as follows:

Compute the shortest path between each pair of vertices (s, t). Determine the shortest path fraction that go through by the vertex V in question. Sum this over all fractions for the pair of vertices (s, t). The betweenness can be more compactly represented as:

$$C_B(v) = \Sigma_{s,t \epsilon V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{5.1}$$

Where (s,t) is the total number of shortest path from node s to t and (s, t| v) is the number of paths that go through v. Betweenness centrality calculates the number of times a node fall on the shortest path among other nodes. Taking graph as input betweenness centrality has been calculated among which the following results take place. Table 5.1 represents the top five hashtags that are the most raised topic among the people in malware #tag.

Table 5.2 represents the most influenced person in the twitter graph i.e. the important nodes which are maximum used for the shortest path to communicate with other followers. Users that remains mostly active in the conversation.

Rank	#tag	Betweenness Centrality values
1.	#cybersecurity	1036.51
2.	#Malware	642.60
3.	#infosec	301.17
4.	#Security	260.41
5.	#ransomware	188.73

Table 5.1: Hashtag Betweenness Centrality values

Rank	From_user screen_name	Betweenness Centrality values
1.	sectest9	517.57
2.	Ovidiug	483.27
3.	securiteIT	280.69
4.	MalwareBeacon	232.97
5.	Evanderburg	181.29

Table 5.2: Screen_Names Betweenness Centrality values

Table 5.3 represents the nodes which get maximum mention by using @tag in the twitter network.

Rank	To_user(@)	Betweenness Centrality values
1.	@josephfcox	212.42
2.	@forenslut	64.95
3.	@CheckPointSW	56.78
4.	@binitamshah	32.59
5.	@DigitalPCSO	28.39

Table 5.3: Mentions Screen_Names Betweenness Centrality values

Figure 5.2 network graph highlights the most influenced people and most discussed topic at the #tag malware in the twitter network. In this Graph network node_color and node_size changes according to betweenness centrality values calculated using Python script. Larger the value of betweenness centrality larger is the size of node in the network which give us a clear view about who talk to whom about what. It gives us the most discussed topic and most influenced person in the network which will be having high betweenness centrality value. Topic will be identifying by #tag in the graph. Table 5.4 represents the overview results of the high value betweenness result of 'Who talks to Whom about What?'

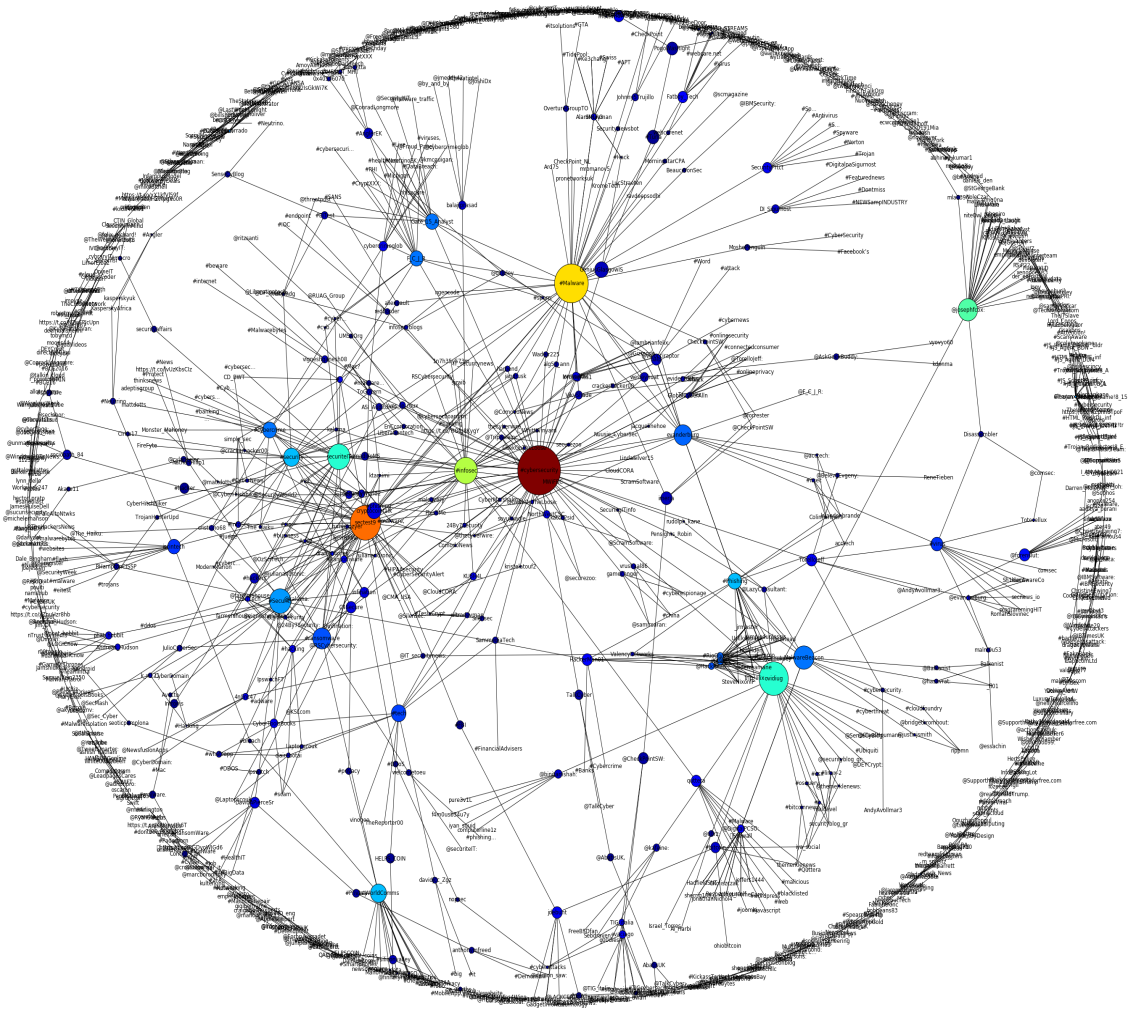

Figure 5.2: Betweenness Centrality graph

So, this result finally leaves us by introducing number of case study. Among which some are explained below:

- **Case Study 1: Background Identity by virtue of posture checking**

  Spying on posture checking of the individuals gives the Background identity that can highly depict their area of interest that can be mostly helpful in hiring job professionals.

  Let A be any cybersecurity company hiring security professionals. Let B, C be the two individuals applying for the job in the cybersecurity company. A is

From_user	About	To_user
Sectest9	#josephfcox #Malware #CyberNews #Infosec #Cybercrime #Security #Ransomware #phishing #CyberSecurityAlert	@CMA_USA @ComodoNews @RSCyberSecurity @infination @CyberHitchhiker
Ovidiug	#bitcoinnews #scams #phishing #Alert #osquery #Cyberthreat #Cybersecurity	@HackRead @Securityblog @binitamshah @senseCyBlog @DEYCrypt

Table 5.4: Who talks to Whom, about What?

in highly search of those people who are having a huge interest in cutting out malware from its root and have a requirement of only one person in their company. Both B, C have pass their criteria, but due to demand of company only one has to get the job. This complex situation can be easily solved by checking their social profiles, which can conclude that 'who talks to whom about what'. It can easily depict their area of interest. From both the person, the one who have talk more about the 'malware', 'security'and many co-related terms will be more desirable to get job in A company. Thus make an easier hand for recruiting job professionals to their area of interest field.

- **Case Study 2: Fake profile identification**

  Inspection of fake profiles by moving through #tags topics that are commonly used between the individuals.

  This type of case study is mostly useful for the security agencies, who are working for the continual check on the activities of criminals and terrorists via the online social networks. There is a person A, who is talking a lot about #tag

'jihad'on the social media. The securities agency is checking A regular updates every day, what A want to do using such 'jihad'#tag. Suddenly A get mute, there is no more tweets or retweets from this profile person but on the same time, someone other called B profile get active to talk about the same hashtag i.e. 'jihad'. They talk a lot and lot about the same topic with the same links as of the previous A. Thus security agencies can conclude that this person B is the same fake profile of the person A.

- **Case Study 3: Defaming personalities**

  Identifying of person trying to defame any personality by using certain #tags in there tweets. This case study will help in catching the criminal activities related to defaming a person using #tag of wrong or vulgar words with mentions name of the personality. This study can clearly depict the idea of who is talking wrong about whom, thus help in identification of person trying to defame the personality.

5.2 Famous personality among Retweets

In the directed graph the edges are created between the from_user screen_name and there to_user screen_name of the retweeted tweets. A directed graph is a graph containing set of edges with direction between the vertices. From each particular tweet edges are created from_user who is sending the tweet to all the to_user who mention in the tweets.

Degree centrality counts the number of links owned by each node. Since, node degree is equal to number of peripheral edges the node has. It tells the number of connections each node is having with rest of the nodes in the network. Degree Centrality for vertex v, is the number of vertices adjacent to it in Graph G = (V, E). For comparison between different graphs, we use the normalized degree centrality:

$$C_D(v) = \frac{deg(v)}{n-1} \tag{5.2}$$

where d is the degree of vertex v, and n the number of vertices of graph G. The higher degree centrality may mean more connections one individual has, thus indicating a larger social circle in a social network. In Directed graph directed edges and its op-

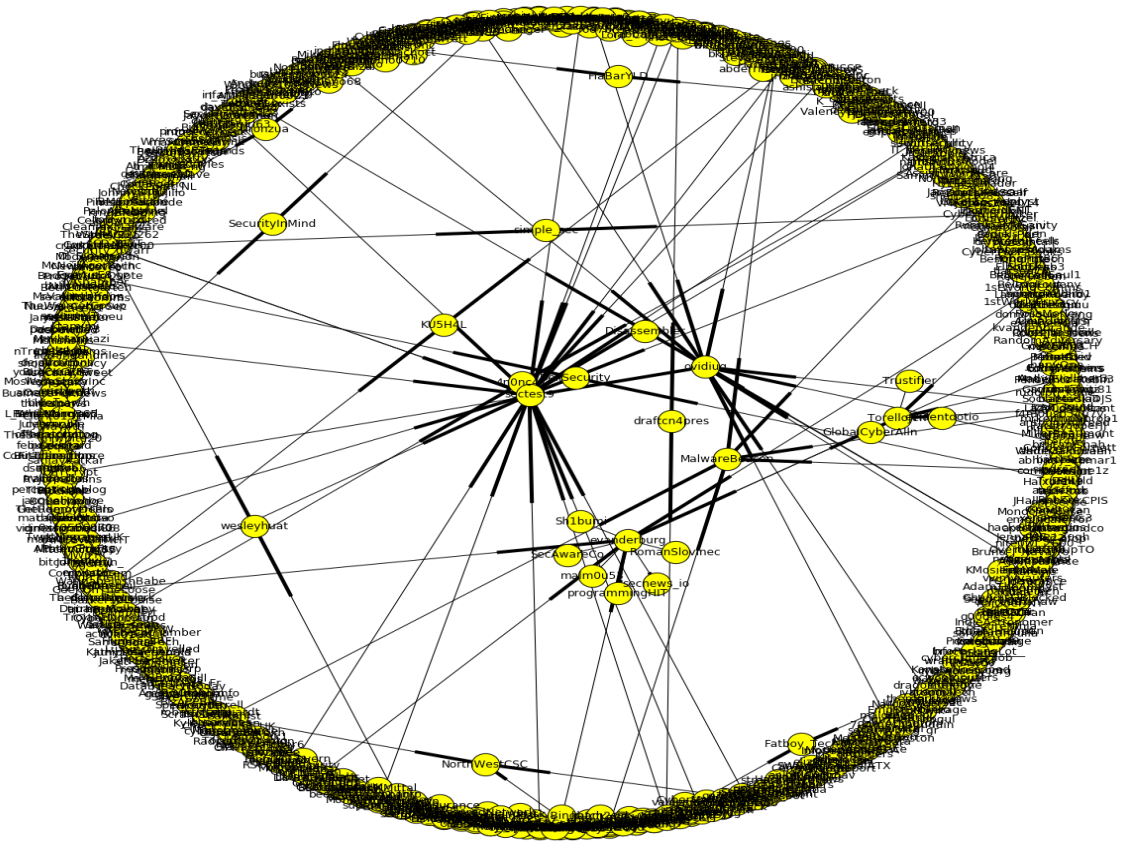

Figure 5.3: Directed Graph for Retweets

erations is common to that of a subclass of graph. There are two types of centrality, Indegree and Outdegree centrality. Indegree centrality gets the value of total number of incoming edges to specific node which symbol the interest towards the nodes whereas Outdegree centrality gets the value of total number of outgoing edges from a specific node and it symbols the activity of the node. Table 5.5 represents the top five In_degree values in the network.

Figure 5.4 represents the network graph that highlights the most influenced people at the #tag malware in the twitter network who make maximum replies to the people in the network. In this Graph node_color and node_size changes according to

Indegree centrality values calculated using Python script. Larger the value of Indegree centrality larger is the size of node in the network which give us a clear view about 'The most influenced speaker'.

Table 5.6 represents the top five In_degree values in the network. Figure 5.5

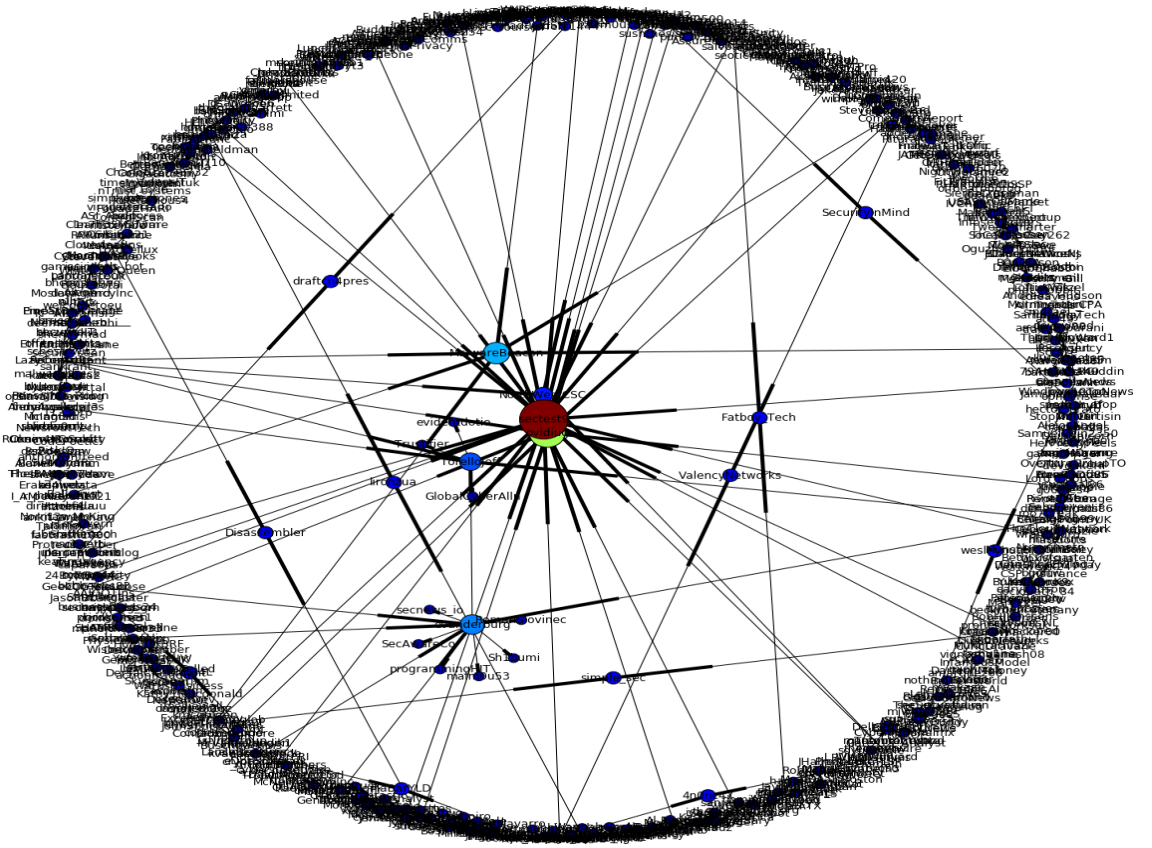

Figure 5.4: In_degree graph

Rank	Screen_name	In_degree values
1.	Sectest9	236.68
2.	Ovidiug	130.17
3.	MalwareBeacon	71.00
4.	Evanderburg	59.17
5.	TorelloJeff	47.33

Table 5.5: In_degree values of retweet graph

represents the network graph that highlights the most influenced people at the #tag

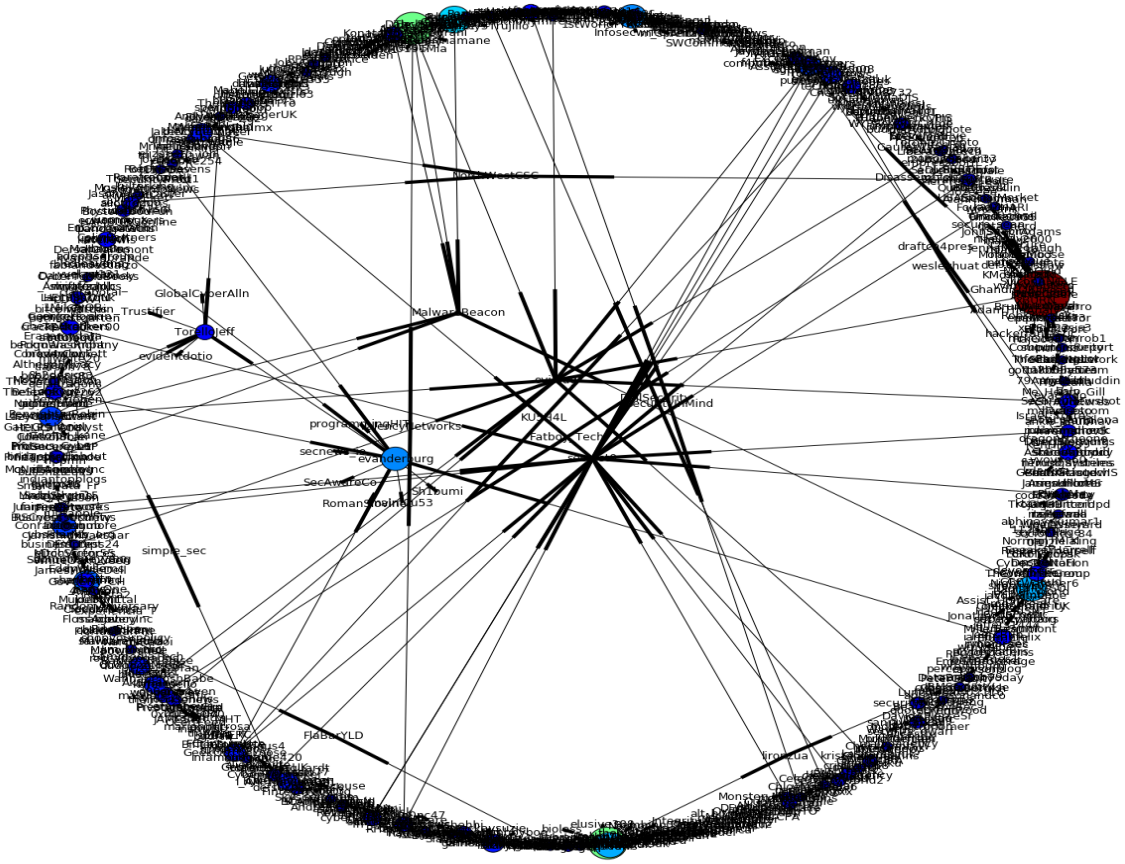

Figure 5.5: Out_degree graph

malware in the twitter network whose tweet get maximum retweeted by the people in the network. In this Graph node_color and node_size changes according to Outdegree centrality values calculated using Python script. Larger the value of Outdegree centrality larger is the size of node in the network which give us a clear view about 'The most retweeted person'.

Rank	Screen_name	Out_degree values
1.	josephfcox	319.52
2.	HackRead	153.84
3.	Hdmoore	153.84
4.	Forenslut	106.50
5.	Actionfrauduk	94.67

Table 5.6: Out_degree values of retweet graph

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this work, a new framework is proposed highlighting use of Centrality. As Centrality indices are important tool that are described on graph vertices to analyze the social network. They are built to interpret the prominence of users according to their position in the network. Python script was implemented for each phase of analyzing tweets in the twitter network. In this Framework NetworkX library in python is used to find the betweenness centrality of the undirected network graph to find 'who talk to whom about what'which pay help in many of the case studies. Also, degree centrality of the directed network graph to find 'most retweeted person'by max Outdegree value and the 'most influence speaker'who make maximum reply to other by max Indegree value.

6.2 Future Work

This study leaves the framework for future studies in several area by python script which would include more detail study about the centrality coming cross to closeness centrality to measure the interaction of people in the network among themselves, which give a brief way of how close relation are there in network.

Further analysis could move to create the visualization of groups in the network based upon the topic about which they discussed more in the network, which may also include the appearance of colored edge by studying more about NetworkX in the network.

References

[1] Brandes, Ulrik. “A faster algorithm for betweenness centrality*.”Journal of mathematical sociology 25.2 (2001): 163-177.

[2] Fisher, Danyel, Marc Smith, and Howard T. Welser. ”You are who you talk to: Detecting roles in usenet newsgroups.” Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS’06). Vol. 3. IEEE, 2006.

[3] Smith, Marc A., et al. “Analyzing (social media) networks with NodeXL.”Proceedings of the fourth international conference on Communities and technologies. ACM, 2009.

[4] Boyd, Danah, Scott Golder, and Gilad Lotan. “Tweet, tweet, retweet: Conversational aspects of retweeting on twitter.”System Sciences (HICSS), 2010 43rd Hawaii International Conference on. IEEE, 2010.

[5] H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a social net-work or a news media? ”in proceedings of 19 international conference on WWW, USA, 2010, pp. 591-600.

[6] M. Cha, H. Haddadi, F. Benevenuto, and K. P.Gummadi, “Measuring user influence on twitter:the million follower fallacy,” in proceedings of 14 Int AAAI Conference on Weblogs and Social Media, Washington DC, 2010, pp. 10-17.

[7] Ediger, David, et al. “Massive social network analysis: Mining twitter for social good.” Parallel Processing (ICPP), 2010 39th International Conference on. IEEE, 2010.

[8] M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," in IEEE conference in social computing, Boston, October 2011, pp. 192-199.

[9] Mincer, Marcin, and Ewa Niewiadomska-Szynkiewicz. "Application of social network analysis to the investigation of interpersonal connections." Journal of Telecommunications and Information Technology (2012): 83-91.

[10] Ellwardt, Lea, Giuseppe Joe Labianca, and Rafael Wittek. "Who are the objects of positive and negative gossip at work?: A social network perspective on workplace gossip." Social Networks 34.2 (2012): 193-205.

[11] Sharma, Naveen Kumar, et al. "Inferring who-is-who in the Twitter social network." ACM SIGCOMM Computer Communication Review 42.4 (2012): 533-538.

[12] Mehmood, Rashid, Helmut Maurer, and Muhammad Tanvir Afzal. "Knowledge discovery in hashtags #." Emerging Technologies (ICET), 2013 IEEE 9th International Conference on. IEEE, 2013.

[13] Lewis, Seth C., Rodrigo Zamith, and Alfred Hermida. "Content analysis in an era of big data: A hybrid approach to computational and manual methods." Journal of Broadcasting & Electronic Media 57.1 (2013): 34-52.

[14] Himelboim, Itai, Stephen McCreery, and Marc Smith. "Birds of a feather tweet together: Integrating network and content analyses to examine crossideology exposure on Twitter." Journal of ComputerMediated Communication 18.2 (2013): 40-60.

[15] Malinick, Todd E., David B. Tindall, and Mario Diani. "Network centrality and social movement media coverage: A two-mode network analytic approach." Social Networks 35.2 (2013): 148-158.

[16] Akhtar, Nadeem, Hira Javed, and Geetanjali Sengar. "Analysis of facebook social network." Computational Intelligence and Communication Networks (CICN), 2013 5th International Conference on. IEEE, 2013.

[17] Akhtar, Naheed. “Social network analysis tools.” Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on. IEEE, 2014.

[18] Jussila, Jari, et al. “Visual network analysis of Twitter data for co-organizing conferences: case CMAD 2013.” System Sciences (HICSS), 2014 47th Hawaii International Conference on. IEEE, 2014.

[19] Himelboim, Itai. ”Political television hosts on Twitter: Examining patterns of interconnectivity and self-exposure in Twitter political talk networks.” Journal of Broadcasting & Electronic Media 58.1 (2014): 76-96.

[20] Kurka, David Burth, Alan Godoy, and Fernando J. Von Zuben. “Online Social Network Analysis: A Survey of Research Applications in Computer Science.” arXiv preprint arXiv:1504.05655 (2015).

[21] Elizabeth I. Peele “Forming your terrorist network: ISIS, Twitter, and the Terrorist propaganda campaign https://cdr.lib.unc.edu/indexablecontent/uuid:5df593ab-141c-4df6-8bf7-d1b235218385 (2015)

[22] Erdogan, Berrin, Talya N. Bauer, and Jorge Walter. ”Deeds that help and words that hurt: Helping and gossip as moderators of the relationship between leadermember exchange and advice network centrality.” Personnel Psychology 68.1 (2015): 185-214.

[23] Dowell, Nia M., et al. ”Modeling Learners’ Social Centrality and Performance through Language and Discourse.” International Educational Data Mining Society (2015).

[24] Chretien, Katherine C., et al. ”A digital ethnography of medical students who use Twitter for professional development.” Journal of general internal medicine 30.11 (2015): 1673-1680.

[25] Corts Martnez, Atia. “Social Network Analysis and the illusion of gender neutral organisations.” (2012).

[26] Freeman, Linton C. “Some antecedents of social network analysis.” Connections 19.1 (1996): 39-42.

[27] Bonacich, Phillip. “Factoring and weighting approaches to status scores and clique identification.” Journal of Mathematical Sociology 2.1 (1972): 113-120.

[28] Zhou, Li. Information Diffusion on Twitter. Diss. University of Dayton, 2015.

[29] Green, Oded, Robert McColl, and David A. Bader. “A fast algorithm for streaming betweenness centrality.” Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). IEEE, 2012.

[30] About Twitter: https://about.twitter.com/company.

[31] Twitter API documentation: http://dev.twitter.com/doc.

[32] “Centrality”, brynmawr, [Online]. Available: http://cs.brynmawr.edu/Courses/cs380/spring2013/section02/slides/05_Centrality.pdf. [Accessed 19 5 2016].

Publication

Navpreet Kaur, Maninder Singh, V.P Singh "Design And Develop A Framework For Social Network Analysis", *International Conference on Inventive Computation Technologies (ICICT 2016) to be held from August 26-27, 2016 at Coimbatore, Tamilnadu, India*[Accepted],ICICT-PaperID: E-311.

Video Link

Url “https://www.youtube.com/channel/UC8K_pMgV83B82azSEEaQx-w”

Plagiarism Report

design and develop the framework for social networking analysis

ORIGINALITY REPORT

9%	3%	8%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Mehmood, Rizwan, Hermann Maurer, and Muhammad Tanveer Afzal. "Knowledge discovery in hashtags#", 2013 IEEE 9th International Conference on Emerging Technologies (ICET), 2013. Publication	2%
2	Green, Oded, Robert McColl, and David A. Bader. "A Fast Algorithm for Streaming Betweenness Centrality", 2012 International Conference on Privacy Security Risk and Trust and 2012 International Confernece on Social Computing, 2012. Publication	2%
3	"Researchers from National Health System Report Recent Findings in Medical Students (A Digital Ethnog", Health & Medicine Week, Dec 4 2015 Issue Publication	1%
4	en.wikipedia.org Internet Source	1%
5	Marc A. Smith. "Analyzing (social media)	