**A**

**Minor Project Report**

on

**Detecting Influencers in Social Networks**

Submitted for partial fulfillment for the degree of

**Bachelor of Technology**

(Information Technology)

in

Department of Information Technology

by

Rishabh Makhija-169108116

Syed Ali-169108149

Under the Guidance of

Mr. Jaya Krishna

(April-2019)

**SCHOOL OF COMPUTING AND INFORMATION TECHNOLOGY**

**MANIPAL UNIVERSITY JAIPUR**



**CERTIFICATE**

Date: 27th April, 2019

This is to certify that the project titled **Detecting Influencers in Social Networks** is a record of the bonafide work done by **Rishabh Makhija** (169108116) and **Syed Ali** (169108149) submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology (B.Tech) in **Information Technology** of Manipal University Jaipur, during the academic year 2018-19.

**Mr. Jaya Krishna**

*Associate Professor*

*Dept. of Computer Science Engineering*

*Manipal University Jaipur*

**Dr. Pankaj Vyas**

*Associate Professor*

*HOD, Dept. of Information Technology*

*Manipal University Jaipur*

**ABSTRACT**

Social networks, which have almost become part of our daily lives, have established new communication structures and behaviors in society. While citizens and businesses have already extensively used social networks for years, governments continuously increase their interest in the new communication technologies. Sites such as Facebook, Twitter, and LinkedIn provide a mechanism for individuals to come together based on a variety of factors such as existing friendships, common interests, or work. People have discovered how the use of social networks can facilitate communication and the exchange/sharing of thoughts and ideas.

So hence, in every network there would be person who would be more influential, and can influence other people to his/her idea. Hence finding an influential person is very important and helps us to spread information.

In this project we worked on centrality measures, information cascading and machine learning techniques to identify the most influential node in the network. Studying different methods best suitable for the network and understanding how information cascading techniques can be applied.

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table No** | **Table Title** | **Page No** |
|  |  |  |
| 1 | Centrality methods before decomposition | 28 |
| 2 | Centrality measures after decomposition | 28 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No** | **Figure Title** | **Page No** |
|  |  |  |
| 1 | Influencers | 8 |
| 2 | Sigmoid Function | 10 |
| 3 | Decision Tree | 12 |
| 4 | Degree centrality | 13 |
| 5 | K-shell decomposition | 14 |
| 6 | Betweenness centrality | 15 |
| 7 | Closeness centrality | 16 |
| 8 | different types of triads | 17 |
| 9 | weibo dataset visualization | 18 |
| 10 | Logistic Regression | 21 |
| 11 | KNN algorithm using k=5 | 22 |
| 12 | Detection of influencers in 9-11 dataset | 23 |
| 13 | Output for logistic classifier | 23 |
| 14 | Output for KNN with k=5 | 23 |
| 15 | Decision tree as entropy | 24 |
| 16 | Decision tree as gini index | 24 |
| 17 | Random Forest | 25 |
| 18 | Saw Mill Dataset | 26 |
| 19 | Karate Club Dataset | 29 |
| 20 | Initializing [0,2,3,4] | 29 |
| 21 | After cascading | 29 |
| 22 | Rishabh Gantt Chart | 30 |
| 23 | Syed Gantt Chart | 30 |

|  |  |  |
| --- | --- | --- |
| **Contents** | | |
|  | | Page No |
| Abstract | | 3 |
| List Of Figures | | 4 |
| List Of Tables | | 4 |
| Contents | | 5 |
| **Chapter 1** | **INTRODUCTION** | 6 |
| **1.1** | Graph theory | 6 |
| **1.2** | Social Networking | 6 |
| **1.3** | Influencers in social network | 7 |
| **Chapter 2** | **BACKGROUND OVERVIEW** | 8 |
| **2.1** | Triads and Triadic Closure | 8 |
| **2.2** | Machine Learning Techniques | 9 |
| **2.3** | Centrality measures | 12 |
| **2.4** | Information Cascading | 15 |
| **Chapter 3** | **METHODOLOGY** | 16 |
| **3.1** | Triadic Closure Analysis | 16 |
| **3.2** | Spammer Detection using Machine Learning | 16 |
| **3.3** | Detecting Influencers using Machine Learning | 17 |
| **3.4** | Combining Local and Global Centrality Measures | 18 |
| **Chapter 4** | **IMPLEMENTATION AND RESULTS** | 19 |
| **4.1** | Spam Detection on weibo dataset | 19 |
| **4.2** | Triadic Closure Analysis on 9-11 dataset | 21 |
| **4.3** | Detecting more influential nodes in twitter dataset | 22 |
| **4.4** | Combining centrality measures | 25 |
| **4.5** | Cascading in a network | 29 |
| **Chapter 5** | **FUTURE WORK AND CONCLUSION** | 30 |
| **5.1** | Progress Chart | 30 |
| **5.2** | Future Work and Conclusion | 31 |
| References | | 32 |

**1] INTRODUCTION**

Online social networks are becoming a bridge that connects our physical daily life with the online world. For example, as of June, 2017, Facebook enjoys 2.01 billion monthly active users. Twitter has 0.65 billion users, who tweet 1 billion times every five days. These connections produce large volumes of data, and its popularity offers the opportunity to study the behavioral and interaction patterns of the data.

* 1. *Graph Theory*

“A picture speaks a thousand words” and a graph speaks so much more than that. Graphs are mathematical structures used to study pairwise relationships between objects and entities, and can represent data visually. This helps us extract valuable information and allows us to take better data driven decisions based on them. A graph is a pair of sets, *G= (V, E)* where V is the set of vertices and E is the set of edges. Vertices are a set of objects or nodes that are connected together using the edges.

There are mainly four types of graphs, namely, connected, unconnected, directed and undirected graphs. If there exists at least one edge between any two nodes, it’s a connected graph. If there exists at least one node that remains unconnected by any edge, it’s an unconnected graph. If all the edges are represented by arrows, it’s a directed graph and if the edges aren’t represented with arrows, it’s an undirected graph.

Graphs are used in numerous fields where data collected from a network can be analyzed. Similarly, graphs can also be used to extract the most influential people in a social network. Advertising companies and marketers can analyze and estimate the person who can market their products over their large social networks and allow larger amount of users to be influenced, in turn increasing the profits. Graphs provide a better way of dealing with abstract concepts like relationships and interactions, thus forming a natural basis for analyzing relationships on a social network.

* 1. *Social Networking*

Social networking is the use of internet-based social media websites in order to connect with your friends, family, colleagues or even clients. Social networking can be used for different purposes, which may be social, business related, or sometimes a combination of both. In the last decade, a large number of social networking platforms have emerged and some of them have grown to have users larger in number than most countries. Few such social network platforms are Facebook, Twitter, LinkedIn, Instagram and many more. All such platforms have become a huge base for companies to create a client-base and engage customers. Companies aim to increase their brand recognition, loyalty and conversion rates. Social networking helps them to access and be recognizable to more new customers, thereby promoting the brand’s name and content.

Social networking offers a lot of advantages to companies in terms of marketing. Customers may complement the company’s offerings and encourage others to buy the products or services. The more customers are talking about a company on social networking, the more valuable the brand authority becomes. As a brand grows stronger, more sales result. Increased company posts rank the company higher in search engines. Social networking can help establish a brand as legitimate, credible, and trustworthy. However, it can be tough to maintain reports or complaints posted on social networks and all social marketing bears some cost. Thus, a company must ensure they invest their social marketing strategies on the correct social influencers so that it can bear positive results for them.

* 1. *Influencers in a social network*

Influence in any network can be termed as the amount of power required to control the decision of a user based on their position in their network. Social influence occurs when one’s opinions, emotions, or behaviors are affected by others.[1] It is being investigated increasingly since it can help spreading messages widely and quickly in the network. In a real life context, an influencer is a person who is followed by many people and has the power to trigger a change in the community. In an online social network, such a person forms large social space where people follow them in large numbers thereby building more and more relationships and expanding their connections. In current times, influencers drive discussions on a particular topic or brand and are also hired by companies to promote or market the products or services. Thus, a social media influencer is a user who has established credibility in a specific industry, has an access to a large audience and can persuade them to act based on their recommendations. They have a reputation for their knowledge and can create trends and encourage their followers to buy products they promote. Figure 1 shows different kinds of influencers online.

In this paper, we address detection of such influencers that can be extracted from the graphs formed by the networks of a social media. Influencers are however, present in any network that shows the relationships between the users. For example, we take a dataset of the 9/11 terrorist network containing the graph that shows the interactions amongst various members of the network. We create a method to extract the five most influential members of the terrorist network that has been discussed in the upcoming sections.

****

Figure

**2 BACKGROUND OVERVIEW**

We aim to detect the influential nodes in a social network by using a number of concepts, namely triadic closure, centrality methods and various machine learning algorithms. A number of other research papers were referred to obtain the required concepts and resources in our research.

*2.1 Triads and Triadic Closure*

We can study social influence from a structure level, focusing on the triadic structure as is the simplest group structure in social networks as well as the cornerstone for studying network formation.[2][3] A triad is a group of three nodes connected together. Basically, there are two types of triads, open and closed. When direction is taken into consideration, 16 such triads can be formed among three nodes. A triad is a very simple and stable structure. It has many features that can be used to the advantage of social network analysis. Every triad follows the transitive property.[4] This can further be explained with a concept called triadic closure property which states that given three nodes A, B and C, if there exists a tie between A-B and A-C, then there must exist a weak or a strong tie between B-C. This property might not be true for every case in a large and complex network but can be helpful as a simple structure to understand networks and predict relationships.

In a referred paper,[5] the authors aim to study the social influence from the perspective of triadic structures to detect its importance in detecting influence. The data sets used are one from a micro-blogging website, weibo.com and the other is used from Cross-Fire which is a shooting game. The method used here is OLS analysis which is ordinary least squares, also known as linear regression. The introduction of the paper mainly focuses on what triads are and how they can have a role in detecting influence in a social network. In their experiment, they show that there are as many as 30 kinds of triads possible considering that there are positive and negative neighbors and strong and weak relationships between any three users. All the users are identified by their unique userIDs and the action considered was retweeting; whether person A, who follows B will re-post some post that person B had posted. In the Cross-Fire dataset, the action considered was if we make free users play games with paying users, would they be influenced to turn into paying users as well. There were 3 types of feature-sets that were taken into consideration. First were basic features which will have the gender, the followers, the people being followed and the number of posts. Second were neighborhood features which contain information about the neighbors of a user, whether they are positive or negative, strongly or weakly connected. The model is trained using basic features first, and combinations of neighborhood features and triadic features are used. The results of their experiment show that when we add that triadic features in regression techniques, the predictive power increases significantly. Thus, this paper provided enough proof and knowledge from the experiments to believe that the triadic structures in any network can trigger influence among its neighbors and can prove to give out better results when used along with other features in the dataset.

*2.2 Machine Learning Techniques*

Machine learning is an application of artificial intelligence that provides the system with an ability to learn and improve from experience to predict patterns or make better decisions. In the paper referred above,[5] it was observed machine learning can help in classifying the more influential nodes in a network, and given they are combined with the right feature sets can produce good results.

We aim to use various machine learning techniques in order to detect the more influential nodes in a given network. We try to do this by using various machine learning classification techniques. Classification is a technique where we categorize data into classes aiming to identify the correct categories for new data. We compare to see which technique best suits this task.

*2.2.1 Logistic Regression*

Logistic regression is a supervised machine learning classification algorithm that is used to analyze a dataset in which there are one or more independent variables that determine the result. The target variable is categorical, and make we use binomial logistic regression in our research, which means the target values has two values, 0 or 1, “yes” or “no”. It predicts the probability of occurrence of an event by fitting the data into a logistic function. The standard logistic function is called the sigmoid function:

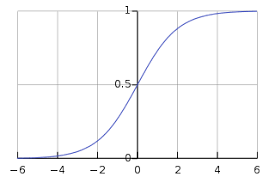


Figure 2- Sigmoid Function

The advantages of using this classifier for our research is that it is very efficient, does not require high computational resources, is highly interpretable, and is easy to regularize. It works better with better feature sets. It is very simple to implement and acts as our base algorithm for comparison with the others. A limitation of this classifier is that it can only predict categorical outcomes. Also, because of its simplicity, it is not the most powerful algorithm and can be easily outperformed by complex algorithms.

*2.2.2 K-Nearest Neighbor*

KNN is another basic yet essential supervised machine learning classification algorithm that is used extensively in pattern recognition and data mining. It can be used to predict both classification and regression problems. The “K” in KNN is the nearest neighbors we wish to take into consideration. The choice of the parameter K is very crucial and it depends on the problem in question. Even with its simple nature, it can give highly competitive results.

Pseudo code for k-NN:  
1. Load the data and initialize the value of k  
2. For getting the predicted class, iterate from 1 to total number of training data points  
 i. Calculate the distance(Euclidean) between test data and each row of training data  
 ii. Sort the calculated distances in ascending order  
 iii. Get top k rows   
 iv. Get the most frequent class from these rows  
 v. Return predicted class

*2.2.3 Decision Tree*

A decision tree is a flowchart-like tree structure where a node represents a feature attribute and the branch represents a decision rule and each leaf node represents the outcome. It partitions the tree recursively on the basis of the attribute values. These values are based on the Attribute Selection Measures (ASM). There are three ASMs, namely Information Gain, Gain Ratio and Gini Index, values of which are calculated for every feature attribute and nodes are selected accordingly.  
ID3 decision tree algorithm uses information gain  
   
   
   
C4.5 uses an extension to information gain known as gain ratio to handle the issue of normalizing information gain using   
   
   
CART decision tree algorithm uses gini method to determine split points

Decision trees can be visualized as a flowchart like structure as shown in Figure 3 that helps to make decisions very easily. Decision trees are capable of handling high dimensional data with good accuracy, although feature selection is crucial to provide best results and prevent overfitting. The decision tree algorithm works on the following basic steps:

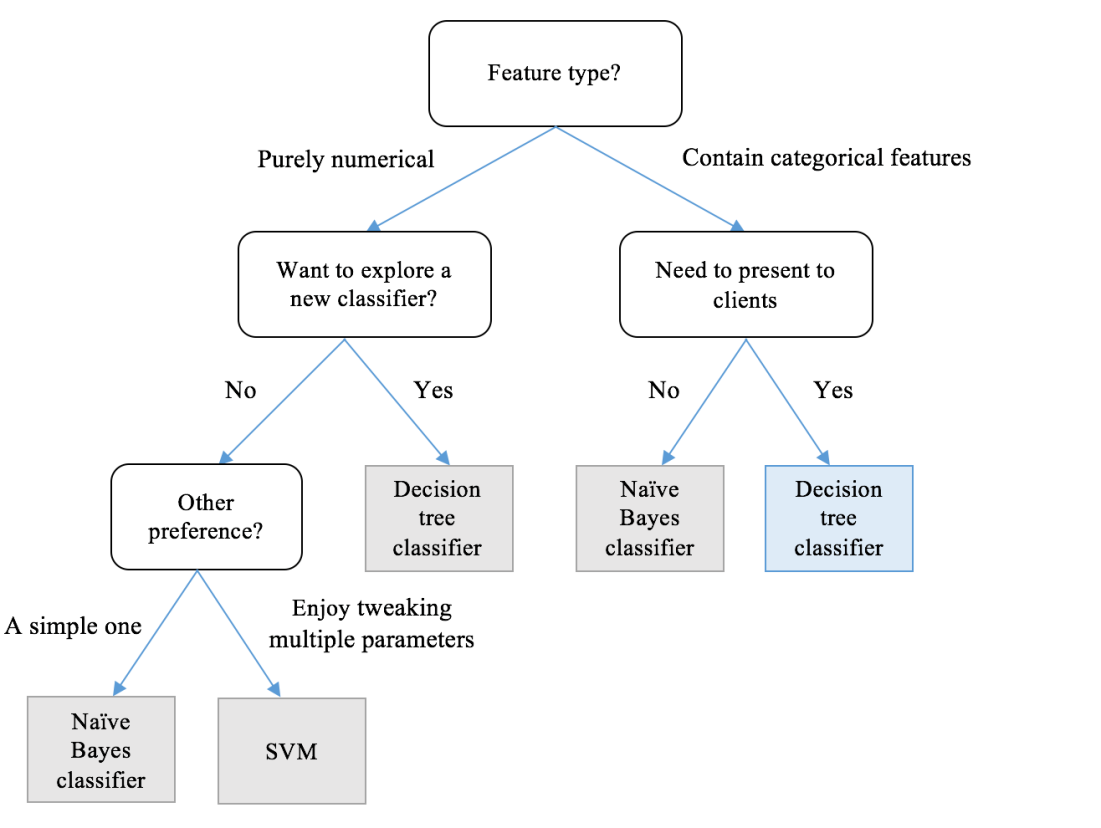
1. Select the best attribute using ASMs  
2. Make that attribute a decision node and divide the data into smaller subsets  
3. Build tree by repeating steps 1 and 2 recursively for each child until either  
 i. All tuples belong to the same attribute; or  
 ii. There are no more remaining attributes; or  
 iii. There are no more instances  


Figure 3- Decision Tree

*2.2.4 Random Forest*

Random forest is a supervised machine learning algorithm that can be used for both classification and regression and is flexible enough to produce great results most of the time. It works by building a “forest” which is an ensemble of multiple random decision trees and merges them together to get a more accurate and stable prediction. Random forest adds additional randomness to the model, while making trees by selecting best features among a random subset of features. At the same time, random forest also provides an easy way to measure the relative importance of each feature. Through looking at the feature importance, we can decide which features to drop to avoid overfitting. A limitation of random forest is that a more accurate prediction requires more trees, which results in a slower model and ineffective for real-time predictions. However, for our experiment, this algorithm seems easy to execute and returns great results too.

*2.3 Centrality*

The term centrality just means the most important part. In graphs and networks it can be referred as the most important nodes. It can also be called as the central nodes of the network. There are different centrality techniques/methods. Centrality methods can be broadly classified into two major classifications.

1. Local centrality measures.
2. Global centrality measures.

*2.3.1] LOCAL CENTRALITY:*

These centrality measures of a node deal with just with the neighboring nodes and not the entire network. This is a fastest method to find the central node but it may not give an accurate solution. In this project we would be dealing with two methods.

1. Degree centrality.
2. K-shell decomposition.

*2.3.2] GLOBAL CENTRALITY:*

Global centrality measures deal with the complete graph. In this method we are finding the central node with respect to the complete network. By this way we are considering the complete graph and the centrality measure finds all the possibilities with the network. In this project we are dealing with two centrality measures.

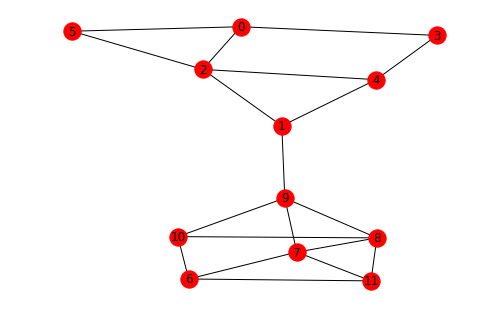
1. Betweenness centrality.
2. Closeness centrality.

*2.3.3] DEGREE CENTRALITY:*

The number of neighboring nodes connected from a node is called as the Degree of the node. Hence, Degree centrality is the ratio of degree of a node by the number of nodes excluding the particular node.

If the network is directed, then the degree centrality is calculated. we have two versions of the measure: in-degree is the number of in-coming links, and out-degree is the number of out-going links. Mostly in a direct graph we would be dealing with in-degree.

Degree Centrality=



Degree of 2: 4

Degree of 5: 2

Figure : Degree centrality

*2.3.4] K-SHELL DECOMPOSITION:*

K-shell decomposition is just the extension of Degree Centrality. For a given graph, they may be nodes with the same degree but depending on the neighbor nodes their influence in the network may vary. Hence, k-shell decomposition identifies the core network and then finds the degree centrality within the core network. This actually involves three steps

**Step 1:** Selecting nodes with degree ‘1’, removing them from the network and adding them to an array.

**Step 2:** Check for nodes with degree ‘1’ after removal and add it the same array.

**Step 3:** Keep on increasing the degree values until all the nodes are classified into a group.

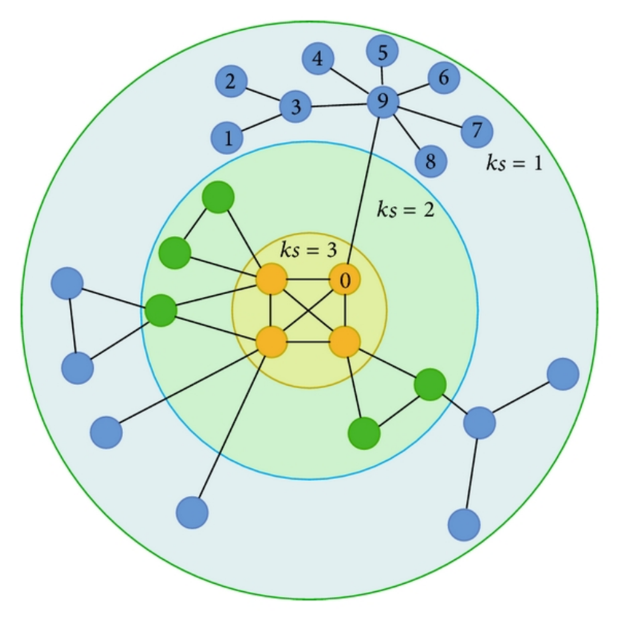


Figure : K-shell decomposition

*2.3.5] BETWEENNESS CENTRALITY:*

Betweenness centrality plays an important role in analysis of social networks, computer networks, and many other types of network data models.

The distance from one node to the other is not the only important factor for information cascading. What is more important is which is the shortest path to for information cascading among the other paths.

Determining betweenness is simple and straightforward when only one node connects each pair of vertices, where the intermediate vertices can completely control communication between pairs of others. But when there are several nodes connecting a pair of vertices, the situation becomes more complicated and the control of the intermediate vertices gets fractionated.

Betweenness Centrality =

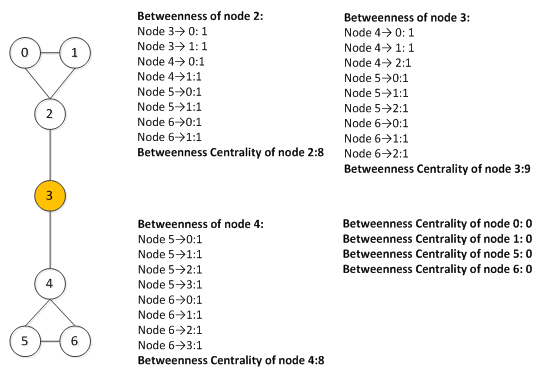


Figure : Betweenness centrality

*2.3.6] CLOSENESS CENTRALITY:*

This Centrality method deals with the closeness between the other nodes, calculated as the reciprocal of the sum of the shortest paths between the node and all other nodes in the graph. The sum gives us the farness hence reciprocal of it gives the closeness. Thus the more central the node is the closer it is to all other node.

The below formula and example is given for an undirected graph. For a directed graph it would totally change depending on the in degree.

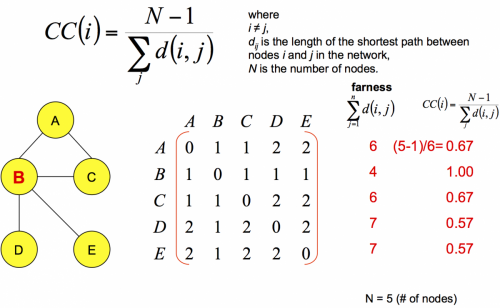


Figure : closeness centrality

*2.4 Information Cascading*

Information cascading is a phenomenon in which a number of people make the same decision. Finding the influential node can help us in effective cascading of information. If we consider a two case scenario, each case might have a payoff load. This is just the weight factor by which that particular task is influential. For example: If a class of 30 students, everyone has to attend a class. But few students plan to go for a movie. Here the payoff load for movie would be more. If the students who plan are the most influential node, then majority of the class would be going for the movie.

**3 METHODOLOGIES**

In order to achieve our aim of detecting influential nodes in a social network and cascading information to a wider range, we use various methodologies as discussed in the coming sub-sections.

*3.1 Triadic Closure Analysis*

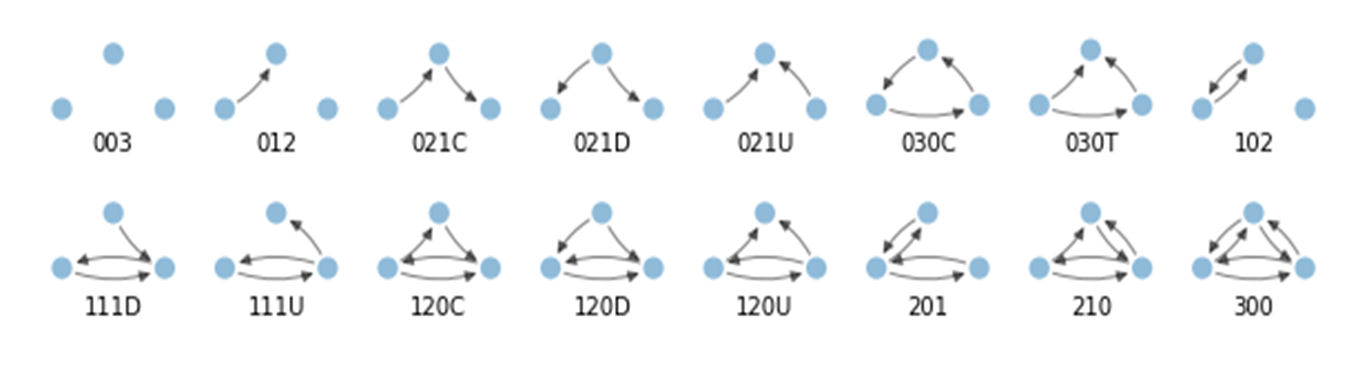
Based on the theory mentioned in section 2.1, we apply the triadic closure analysis on a terrorist network of 9-11. The basic idea behind the methodology is to extract the frequency of each number of triads in the network.[6] There are 16 different types of standardized triads, as shown in Figure 4, that we work with for this experiment and the triadic closure analysis helps us to gather which node has which kinds of triads and how many in amount. From this result, we further use it to form a new table and apply probabilistic formula along with weights assigned to the type of triads, and calculate an influence index for each node in the network. We apply some threshold values to extract the nodes with influence index higher than the threshold value. We sort the results to get the most influential nodes of the network and try to visualize the influential nodes and their reach in the network.   


Figure : different types of triads

*3.2 Spammer Detection using Machine Learning*

Based on the theory mentioned in section 2.2, we explore an application of machine learning classification algorithms known as spam detection. We assume that any spammer is likely to have more followees than followers, as it has been observed by the social networks. The number of posts made by a normal non-spamming user are more in number than a spammer. Spammers are more likely to have a lower class or priority in the network than a normal user. It is observed in this dataset that most of the spammers are in the initial lower classes and that the higher classes have rare or no occurence of spammers. Due to its ever-growing nature, social networking platforms also end up holding a lot of spam profiles or abandoned profiles that haven’t yet been removed from their databases. However, focusing on spam profiles, they tend to reflect similar characteristics to that of an influential profile. Thus, to maintain authenticity and increase the level of security in our research, we explore detection of spammers in a dataset taken from a micro-blogging site weibo.com. We apply logistic regression and k-NN algorithm to predict the spammer nodes in the dataset. Both these algorithms are easy to apply and produce good results with great efficiency and speed. Also, since this task does not require a complex algorithm, working with the simpler classification algorithms is preferred.

Figure 9- weibo dataset visualization

*3.3 Detecting Influencers using Machine Learning*

Based on the theory mentioned in section 2.2, we aim to detect which nodes in a network can be more influential than others. Using a twitter dataset, we try to classify between two users, which is more influential than the other based on the feature attributes available. To select the most suitable features, we use a forward-backward stepwise selection based on p-values of co-efficients of linear regression. Next we create a model function that could train, test, validate and present the results of the various algorithms used. When the dataset along with the classifier object of the algorithm is passed into the function, it undergoes the following steps:  
i. Split the dataset into train and test dataset, splitting percentage=30%  
ii. Train and fit the algorithm  
iii. Make predictions on test datset  
iv. Generate evaluation metrics (confusion matrix, accuracy score, classification report)  
v. Draw ROC Curve   
vi. Cross validation analysis based on accuracy score and auc score

We use logistic regression as our base algorithm and implement it by passing the classifier object and dataset to the model function. We next apply k-NN to classify the more influential nodes, while keeping the value of k as 5. Similarly, we apply decision tree classifiers based on splitting criteria as entropy and gini index. We also apply random forest classifier keeping its splitting criteria as entropy. Finally, we make a comparison as to which algorithm fared better in classifying the more influential nodes in the dataset. The results obtained by each algorithm shows the confusion matrix it forms and the corresponding evaluation metrics like accuracy, precision, recall, F1 score as well as the ROC curve, that help us to decide which algorithm classified the problem best.

*3.4 COMBINING LOCAL AND GLOBAL CENTRALITY MEASURES*

We know that global centrality methods are precise and time consuming whereas local centrality gives us a quick result but might not be precise. Hence we come up with an algorithm to combining them [6]. Initially we implement K-shell decomposition in the given network, remove the nodes which doesn’t belong to the core network. This will decrease a significant no. of nodes and we can concentrate on the core network only. Then applying the remaining centrality theorem that is, Degree centrality, betweenness centrality and closeness centrality.

**4 IMPLEMENTATION AND RESULTS**

Implementations of all our experiments have been carried out in python language on Jupyter notebook.

*4.1 Spam Detection on weibo dataset*

We used classification machine learning algorithms, namely logistic regression and k-nearest neighbors to detect whether a user in a dataset is a spammer or not. The dataset has7 attributes. After analyzing the dataset, and using the graphs in Figure 5, we make some observations:  
i. Spammers are more likely to have a lower class than a normal user. The highest frequency of spammers occurs in class 3, followed by class 2 & 1. Therefore, the number of followees in class 3 is higher than any other class. There are no spammers present in class 10 or above.

ii. Spammer profiles are higher in frequency with the female gender than male gender. This mildly suggests that males create female profiles in order to spam posts.

iii. It is observed here that highest number of users are present in class 7-8, second highest in class 3-4. The users are in the initial lower classes have spammers present and that the higher classes have rare or no occurrence of spammers.

We select gender, number of posts, number of followers and the class the user belongs to as the feature set. We then scale the dataset and transform them. We apply logistic and KNN algorithms and compare the results as shown in Figure 6 and 7 below.

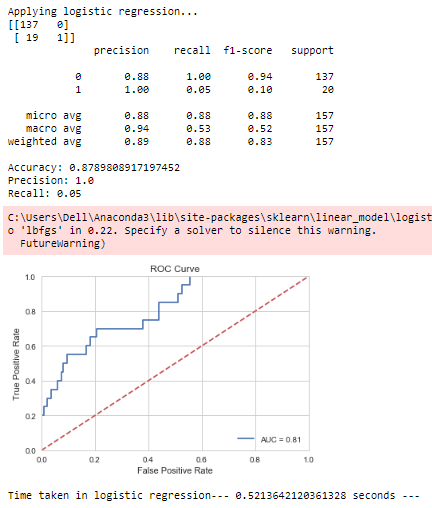


Figure 10- Logistic Regression

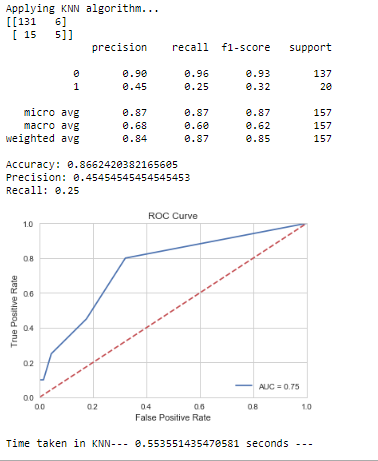
**

Figure 11- KNN algorithm using k=5

*4.2 Triadic Closure Analysis on 9-11 dataset*

We implemented triadic closure analysis on a data of a terrorist network of 9-11. The dataset had an edge-list with weights assigned based on the level of interactions the nodes had. A second file contained the ‘flight’ attribute that denotes to which flight was the node related to. We perform the triadic closure on the all the nodes that had the ‘flight’ attribute and calculate the frequencies of each the 16 triads present in Figure 4. It is observed that three types of triads, namely 102, 201 and 300 occur in maximum frequencies and an influence-index is calculated using the formula below:

Using this formula while keeping a threshold as 0.2 and calculating the influence index for each node in the network, we can detect the influencers in the network. Figure 8 shows the five most influential nodes extracted from the network.

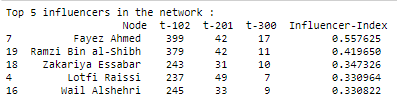


Figure 12- detection of influencers in 9-11 dataset

*4.3 Detecting more influential nodes in twitter dataset*

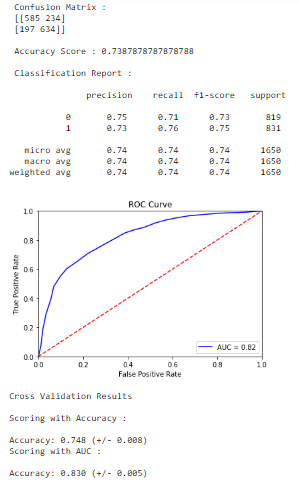
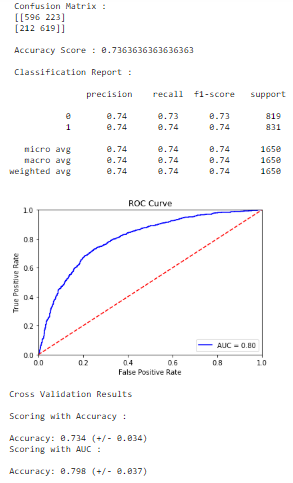
We made use of various machine learning techniques to detect the more influential node among two users, given in a large twitter dataset. The dataset had 23 attributes. On using the co-relation heatmap, we could observe which features had high co-relation. We applied a forward –backward feature selection method based on the p-values. Using the resultant feature set, we applied the following machine learning classifiers available on python’s sklearn library, and the results for the same have been shown along with figures below:  
i. Logistic regression ii. KNN algorithm  


Figure -Output for logistic classifier Figure - Output for KNN with k=5

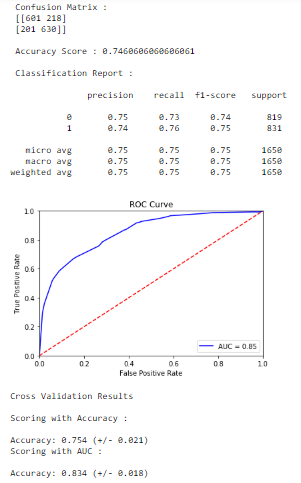
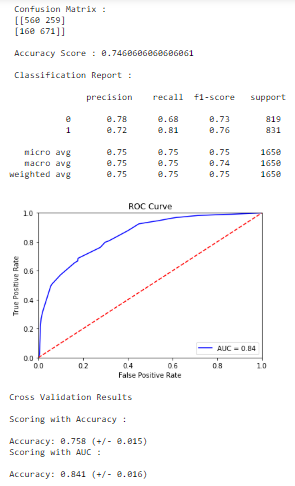
iii. Decision tree keeping splitting criteria as entropy and gini index respectively  


Figure - Decision tree as entropy Figure - Decision tree as gini index

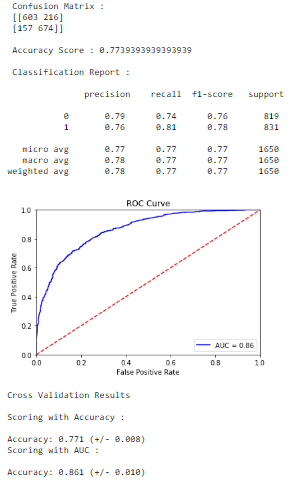
iv. Random forest  


Figure - Random Forest

Among the classification results produced by these five algorithms, random forest outperforms them all. Even though the results don’t indicate significant differences, but in a large dataset, the minutest of difference in performances can cause a great deal. However, random forest does take considerably more amount of processing time as compared to other classifiers, but since our problem statement does not require a real time analysis of data, random forest turns out to be the better suited algorithm for detecting and classifying the more influential nodes in a dataset.

*4.4 Combining centrality measures*

To combine centrality measures we are implementing this on the well-known dataset (Saw mill). The graph representing it is shown below.

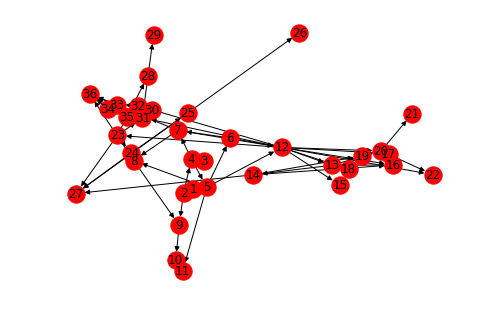


Figure - Saw Mill Dataset

Before implementing k-shell decomposition the centrality results are shown in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Before Removing | | |
| Nodes | Degree | betweeness centrality | Closeness Centrality |
| 1 | 1 | 0 | 0 |
| 2 | 3 | 0 | 0.028571429 |
| 3 | 1 | 0 | 0 |
| 4 | 4 | 0.083193277 | 0.064285714 |
| 5 | 5 | 0.039327731 | 0.057142857 |
| 6 | 3 | 0 | 0.054945055 |
| 7 | 5 | 0.037815126 | 0.093506493 |
| 8 | 4 | 0 | 0.0875 |
| 9 | 3 | 0.004537815 | 0.101587302 |
| 10 | 2 | 0.004537815 | 0.085714285 |
| 11 | 2 | 0 | 0.102040816 |
| 12 | 13 | 0.063529412 | 0.093333333 |
| 13 | 4 | 0.00605042 | 0.079503106 |
| 14 | 4 | 0 | 0.072321429 |
| 15 | 2 | 0 | 0.079503106 |
| 16 | 4 | 0.00302521 | 0.114285714 |
| 17 | 3 | 0.00907563 | 0.123469388 |
| 18 | 3 | 0.00605042 | 0.096428571 |

|  |  |  |  |
| --- | --- | --- | --- |
| 19 | 3 | 0 | 0.111196911 |
| 20 | 6 | 0.01210084 | 0.183673469 |
| 21 | 1 | 0 | 0.143417367 |
| 22 | 2 | 0 | 0.170099668 |
| 23 | 5 | 0.036302521 | 0.079503105 |
| 24 | 3 | 0.01210084 | 0.072321429 |
| 25 | 3 | 0.00605042 | 0.068027211 |
| 26 | 1 | 0 | 0.06522911 |
| 27 | 4 | 0 | 0.130501931 |
| 28 | 3 | 0.00605042 | 0.072321429 |
| 29 | 1 | 0 | 0.068027211 |
| 30 | 2 | 0 | 0 |
| 31 | 7 | 0.011092437 | 0.132967033 |
| 32 | 4 | 0.004033613 | 0.141871921 |
| 33 | 3 | 0.003529412 | 0.123809524 |
| 34 | 2 | 0 | 0 |
| 35 | 3 | 0.002521008 | 0.123809524 |
| 36 | 5 | 0 | 0.223166023 |

Table 1- Centrality methods before decomposition.

From this table their exist a unique node for each centrality measure.

After performing K-shell decomposition and removing the unrelated nodes. The table is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  | After K-shell Removal | | |
| Nodes | degree | betweenness centrality | closeness centrality |
| 5 | 3 | 0 | 0 |
| 6 | 3 | 0.009848485 | 0.083333333 |
| 7 | 4 | 0.019696969 | 0.111111111 |
| 8 | 3 | 0.009848485 | 0.1875 |
| 12 | 9 | 0.206818182 | 0.25 |
| 13 | 3 | 0 | 0.19047619 |
| 16 | 3 | 0 | 0.260416667 |
| 17 | 3 | 0.009848484 | 0.3 |
| 20 | 3 | 0 | 0.371212121 |
| 31 | 3 | 0.024621212 | 0.19047619 |
| 32 | 3 | 0.024621212 | 0.260416667 |
| 33 | 3 | 0.01969697 | 0.23076923 |
| 36 | 3 | 0 | 0.380952381 |

Table 2- centrality measures after decomposition(note: other nodes are removed)

The results are as expected since we got a common centrality node i.e, 12 and the difference between the closeness centrality of 36 and 12 is minimum and hence we can consider it.

*4.5 Cascading in a network*

For understanding cascading, we would be requiring a directed graph. For this we are making use of a very common undirected graph, i.e. The Karate club. We have just considered two different activities taking place in the network. Action A and Action B. We assume that initially the complete network is performing activity ‘B’. Let the payoff(A) = a= 6 and payoff(B)=b=3. By self-inputting few nodes to be working on Activity A. Influence in the node is shown below.

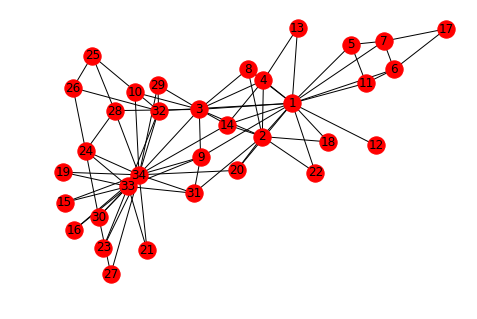
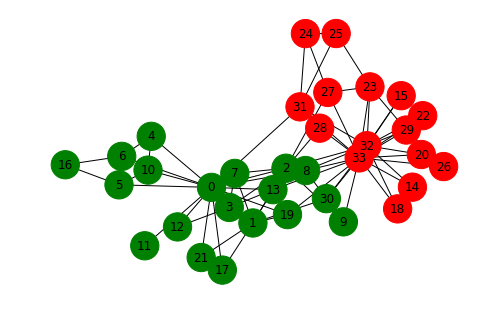


Figure - Karate Club

Initializing with nodes [0, 2, 3, 4]



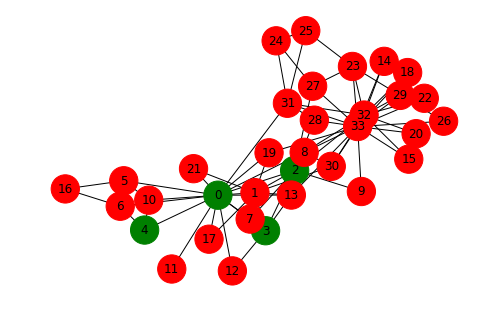


Figure - initializing [0,2,3,4] Figure - after cascading

**5 FUTURE WORK AND CONCLUSION**

*5.1 Progress chat*

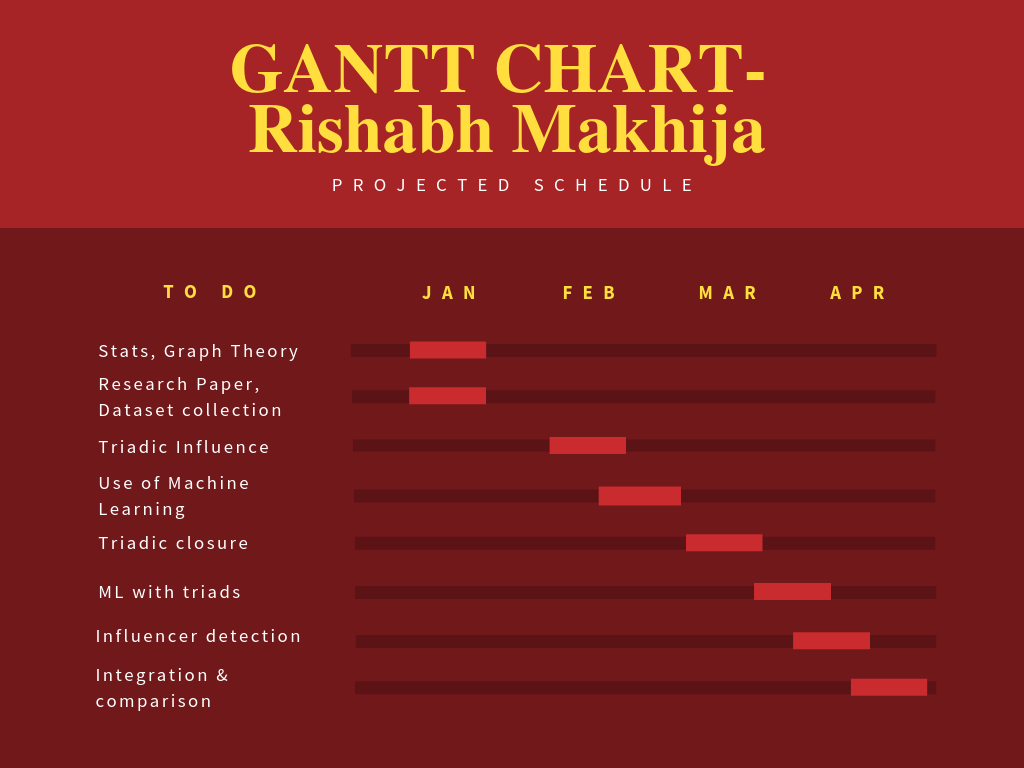
**

Figure - Rishabh Gantt chart

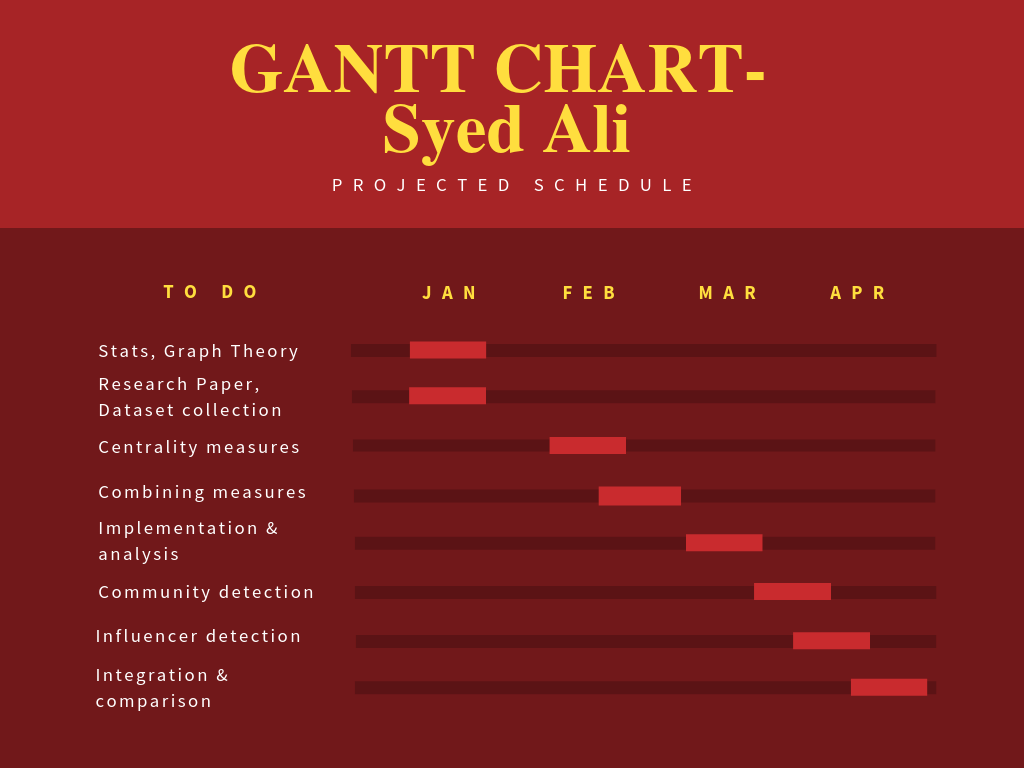


Figure - Syed Gantt Chart

*5.2 Future work and conclusion:*

In terms of related work in the future, we identify the following areas:  
i. Use of parallelization to try and reduce the computation time.   
ii. Evolution of possible generic laws to govern the formation of networks.

iii. work on the centrality methods and understand cascading after k-shell decomposition.

In our research, we explore various methods that can be used to detect influencers in a social network. While implementing triadic closure on a network that is not typically social, and using a probabilistic factor for calculating metrics, we extract the influencers with large interconnectivity in the network. We also use various machine learning techniques to achieve our aim of detecting influencers, and in a comparative analysis conclude that random forest is best suited when it comes to classifying the dataset of a social network that is full of people engaging in random actions. Our experimental results confirm that random forest is a great choice for solving such a classification problem, especially in terms of performance.

We also worked on the centrality methods and ways to combine different centrality methods to get an optimum result. And also worked on a cascading method to understand how cascading works in a network, how influential nodes affect the decision and how payoff affect the influence.

**REFERENCES:**

*Journal / Conference Papers*

1. M. Deutsch & H. B. Gerard, “A study of normative and informational social influences upon individual judgment,” Journal of Abnormal and Social Psychology, vol. 51, pp. 629–636, 1955.
2. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii & U. Alon, “Network motifs: simple building blocks of complex networks,” Science, pp. 824–827, 2002.
3. H. Huang, J. Tang, L. Liu, J. Luo & X. Fu, “Triadic closure pattern analysis and prediction in social networks,” IEEE TKDE, vol. 27, pp. 3374–3389, 2015.
4. D. Easley & J. Kleinberg(2010). “Networks, crowds and markets: reasoning about a highly connected world.” Cornell, NY: Cambridge Univ Pr.
5. C. Zhang, S. Gao, J. Tang(2016). “Learning triadic influence in large social networks.” IEEE/ACM International Conference on ASONAM.
6. Xia, Y., Ren, X., Peng, Z.(2014). “Effectively identifying the influential spreaders in large-scale social networks”.

.

*Web*

[1] Social Networks, NPTEL.