A Study on Influence Maximization Problem in Social Network

Submitted by:

Souvik Saha

T91/IT/154004

D01-1121-0171-15

Soumyajit Bhattacharyya

T91/CSE/156001

A03-1112-0120-13

Pratik Mukherjee

T91/CSE/154013

D01-1121-0127-15

In partial fulfilment for the Award of the Degree of Bachelor of Technology in Computer Science and Engineering, University of Calcutta

Under the supervision of:

Dr. Sankhayan Choudhury

Professor, Department of Computer Science and Engineering

University of Calcutta

University of Calcutta

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the project entitled "***A Study on Influence Maximization Problem in Social Network*** " is a bonafide record of the work done by Souvik Saha(T91/IT/154004), Soumyajit Bhattacharyya (T91/CSE/156001) and Pratik Mukherjee(T91/CSE/154013) under the supervision of Dr. Sankhayan Choudhury , in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Computer Science And Engineering from Department of Computer Science And Engineering , University of Calcutta for the session 2018-2019.

Dr. Banani Saha
Chairman, Board of Studies,
Computer Science &Engineering
University of Calcutta

Dr.Sankhayan Choudhury
Professor, Department of Computer Science & Engineering
University of Calcutta

External Examiner

ACKNOWLEDGEMENT

We have received the assistance and cooperation of quiet a number of people during our project, ***A Study on Influence Maximization Problem in Social Network***. So, we hereby take the opportunity to extend our sincere gratitude to all those who have provided their assistance and cooperation and valuable suggestion from time to time in spite of their busy schedule. First, we would like to thank our project guide Prof. Sankhayan Choudhury. He has not just helped us but provided his valuable and insightful suggestion from time to time for generating in us a profound interest for this subject that kept us motivated during the entire duration of this project. Also, we thank Shoubhik Goswami, Faculty Member, Meghnad Saha College of Engineering, Kolkata, for his valuable advice and inputs.

We would also like to extend our sincere gratitude to Prof Banani Saha and entire Technical Assistant team of Department of CSE for their encouraging support and for providing us a pleasant atmosphere during the period of project development. Lastly, we would like to thank our parents for their overall support.

Souvik Saha

Roll No.-T91/IT/154004

Regn. No.-D01-1121-0171-15

Soumyajit Bhattacharyya

Roll No.-T91/CSE/156001

Regn. No.- A03-1112-0120-13

Pratik Mukherjee

Roll No.-T91/CSE/154013

Regn. No.- D01-1121-0127-15

Contents

Chapter 1

Introduction

1.1 Overview

Finding the most influential people is an NP-Hard problem that has attracted many researchers in the field of social networks. The problem is also known as influence maximization and aims to find the number of people that are able to maximize the spread of influence within a target social network. The influence problem is motivated by many applications in different fields: from marketing, with the aim of maximizing the adoption of a new product. The huge structure of the social networks poses a great problem for researchers to find an efficient heuristic solution of the problem. Our project focusses on an experiment to measure and compare different metrics to maximize the influence on the network by finding communities and selecting specific nodes to spread the influence.

1.2 Broad Aspect and Applications

Influential nodes are rare in social networks, but their influence can quickly spread to most nodes in the network. Identifying influential nodes allows us to better control epidemic outbreaks, accelerate information propagation, and conduct successful e-commerce advertisements, and so on. Classic methods for ranking influential nodes have limitations because they ignore the impact of the topology of neighbour nodes on a node.

The increasing popularity of many online social network sites, such as Facebook, Myspace, and Twitter, presents new opportunities for enabling large-scale and prevalent viral marketing online. Consider the following hypothetical scenario as a motivating example. A small company develops a cool online application for an online social network and wants to market it through the same network. It has a limited budget such that it can only select a small number of initial users in the network to use it (by giving them gifts or payments). The company wishes that these initial users would love the application and start influencing their friends on the social network to use it, and their friends would influence their friends' friends and so on, and thus through the word-of-mouth effect a large population in the social network would adopt the application. The problem is whom to select as the initial users so that they eventually influence the largest number of people in the network, i.e., the problem of finding influential individuals in a social network.

In this project different metrics has been used to measure the degree of influence the influence in the network. Degree centrality, Betweenness centrality, Closeness centrality, Eigen vector centrality has been used to measure the importance of a node in a network efficiently. Using these metrics, the most important nodes through which the influence will be diffused in the network can be determined efficiently. By now, personal life has been invaded by online social networks (OSNs) everywhere. They intend to move more and more offline lives to online social networks. Therefore, online social networks can reflect the structure of offline human society. A piece of information can be exchanged or diffused between individuals in social networks. From this diffusion process, lots of latent information can be mined. It can be used for market predicting, rumour controlling, and opinion monitoring among other things.

1.3 Case Study

Recently many large-scale online social network sites, such as Facebook and Friendster, become successful because they are very effective tools in connecting people and bringing small and disconnected offline social networks together. Moreover, they are also becoming a huge dissemination and marketing platform, allowing information and ideas to influence a large population in a short period of time. However, to fully utilize these social networks as marketing and information dissemination platforms, many challenges have to be met.

This problem, referred to as influence maximization, would be of interest to many companies as well as individuals that want to promote their products, services, and innovative ideas through the powerful word-of-mouth effect (or called viral marketing). Online social networks provide good opportunities to address this problem, because they are connecting a huge number of people and they collect a huge amount of information about the social network structures and communication dynamics. From the below examples we can have some ideas of implementation of influence maximization through some influential node.

70% of millennial consumers are influenced by the recommendations of their peers in buying decisions. An influencer marketing survey conducted by Collective Bias involving 14,000 respondents in US, reveals this statistic. The same survey reveals that 30% consumers are more likely to buy a product recommended by a non-celebrity blogger. Consumers can relate more to these influencers and value their opinion more than that of celebrity influencers. Facebook is the most influential social channel per the results of the survey. 19% of consumer purchase decisions are influenced by Facebook posts. Just 3% of consumers are influenced by celebrity endorsements in their product purchase decisions. This influencer marketing stat was also discovered in the Collective Bias survey. YouTube seems to be the second most influential social media platform for purchases. 18% of consumers are influenced by YouTube regarding their purchases.

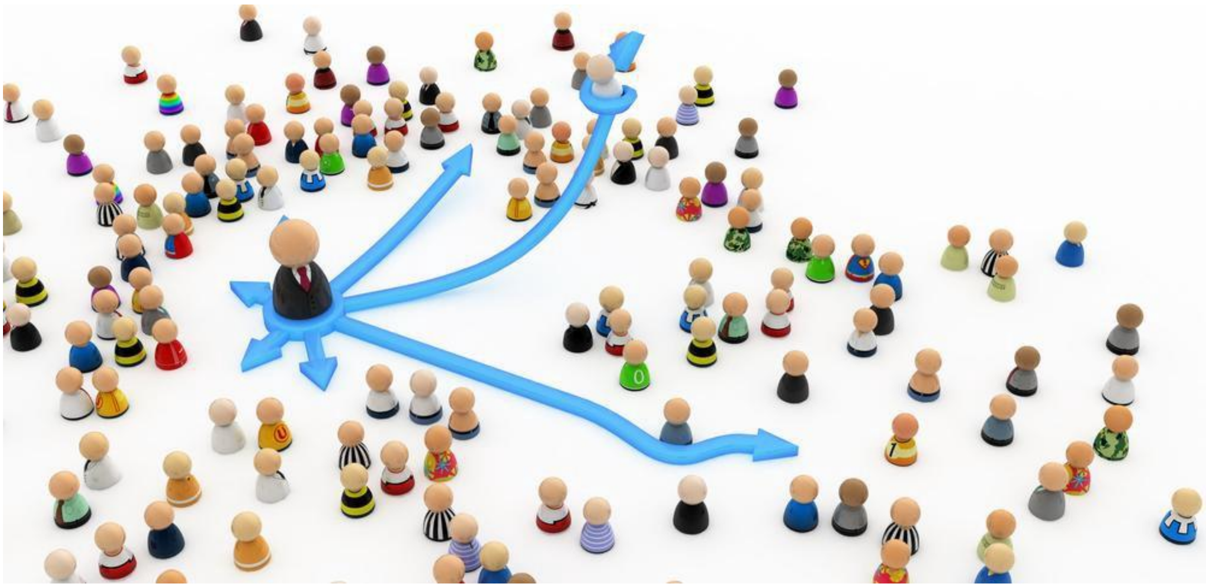


Fig-1 Influence spreading in a network.

The main goal of this project is to conduct a brief study on the "Influence Maximization" problem, including the existing notable solution approaches.

The organization of the dissertation is as follows. Chapter 2 describes the pre-requisites for the subsequent discussion on the said problem. The present state of art is discussed in chapter 3 that includes the problem statement also. The chapter 4 illustrates the experiment details and the findings regarding the selected algorithms. The dissertation concludes in chapter 5.

Chapter 2

Preliminaries

In this chapter we introduce the required concepts of community in a social network, and different measures of influence strength which helps us to establish the methods used in the experiments.

2.1 The Concept of Community

Community is a group of nodes that have some common properties and have common role in organization. Group of nodes are more densely connected if they belong to the same community and less likely to be connected if they are not the members of same community. In general, we consider a particular set of nodes within a specified community by considering that there will be more edges within a particular set of nodes (i.e. more intra edges than inter edges) compared to the rest of the nodes in the graph and hence we isolate that group of nodes as a community. Newman calls the group of vertices with dense connections within group and sparser connections between group communities.

Different types of communities are as follows:

i. **Disjoint Community**: In disjoint community [3] a node belongs to single community. It has crisp assignment, where binary relationship is being held between a node and a community. An example is shown in fig-2.

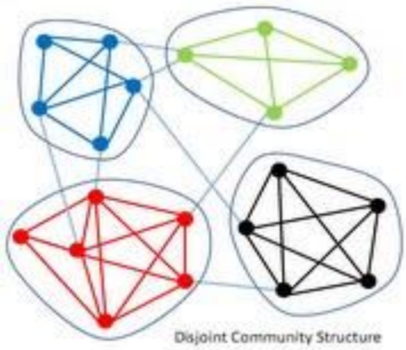


Fig-2

ii. **Overlapping Community**: In overlapping community a node may belongs to more than one community. This is known as fuzzy assignment of nodes where a node may belong to more than one community. An example is shown in fig-3.

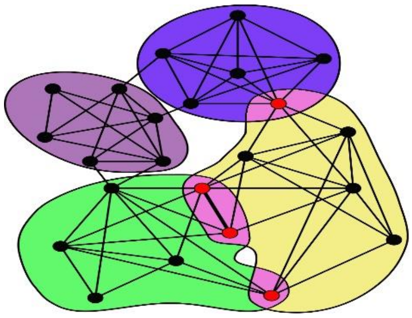


Fig-3

2.2 Basic Measures of Influence Strength

A social network is modelled as a graph G = {V, E}, where V is the set of nodes, and E is the set of edges. As per the convention, the links correspond to actors (people) and the links correspond to social relationships. At the local level, social influence is a directional effect from node A to node B, and is related to the edge strength from A to B. On a global level, some nodes can have intrinsically higher influence than others due to network structure. These global measures are often associated with nodes in the network rather than edges. We are presenting the concepts and measures at a local and global level respectively.

2.2.1 Edge Measures

- ***Tie Strength***

Edge or tie strength concept was introduced by Granovetter. On edge level there are two different types of ties, strong ties and weak ties. The tie depends on the number of overlapping friends or neighbours between two nodes. The larger the overlap the stronger the ties between the nodes. The strength between two nodes A and B can be defined as

$$S(A,B) = \frac{|n_A \cap n_B|}{|n_A \cup n_B|}$$

where n_A and n_B are the neighbours of nodes A and B, respectively.

Strong ties represent trust relationship between nodes or simply friendship, whereas Weak ties occur between acquaintances when the friendship overlap is small and restricted information is shared between the nodes such as private profiles, hidden personal details and private posts. The fig-4 shown below depicts strong tie as whereas fig-5 depicts weak ties. Sometimes, the tie strength is defined under a different name called embeddedness. The embeddedness of an edge is high if two nodes incident on the edge have a high overlap of neighbourhoods. When two individuals are connected by an embedded edge, it makes it easier for them to trust one another, because it is easier to find out dishonest behaviour through mutual friends. On the other end, when embeddedness is zero, two end nodes have no mutual friends. Therefore, it is riskier for them to trust each other because there are no mutual friends for behavioural verification. When the overlap of the neighbourhoods of A and B is small, the connection A-B is considered to be a weak tie. When there is no overlap, the connection A-B is a local bridge. In the extreme case, the removal of A-B may result in the disconnection of the connected component containing A and B. In such a case, the connection A-B may be considered a global bridge.

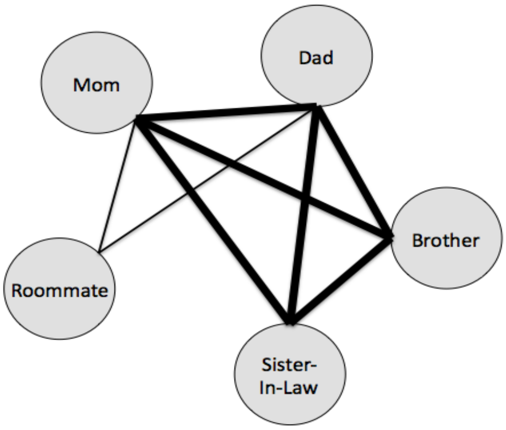


Fig-4 Strong Tie

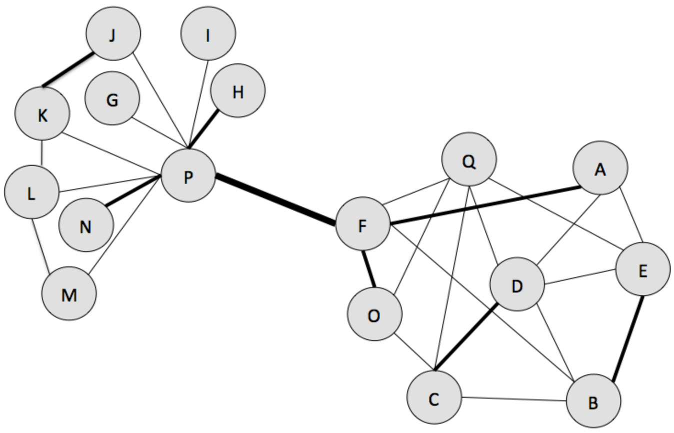


Fig-5 Weak Tie (Bridge)

- ***Edge Betweenness***

The edge betweenness centrality is defined as the number of the shortest paths that go through an edge in a graph or network. Each edge in   can be associated with an edge betweenness centrality value. The edge betweenness centrality value for edge (a, b) is computed from the number of pairs of nodes x and y such that the edge (a, b) lies on the shortest path between x and y. Since there can be several such shortest paths edge (a, b) is credited with the fraction of those shortest paths that include (a, b).

$$\mathrm{bt}(a,b) = \sum_{x,y} \frac{\#shortest\_paths(x,y)through(a,b)}{\#shortest\_paths(x,y)}$$

An example is shown in fig-6 to depict the calculation of edge betweenness.

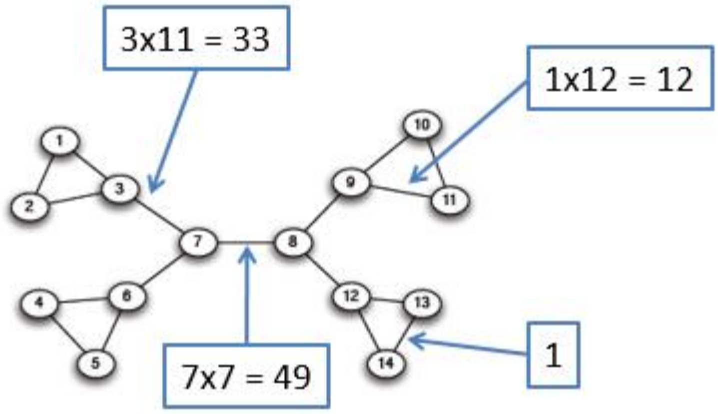


Fig-6

Edges that have a high probability to occur on a randomly chosen shortest path between two randomly chosen nodes have a high betweenness.

The famous Girvan-Newman method is based on edge betweenness centrality.

The steps to find the betweenness is given below-

- Perform a BFS starting from A.
- Determine the shortest path from A to each other node.
- Based on these numbers, determine the amount of flow from A to all other nodes that uses each edge.
- Repeat the process for all nodes.
- Sum over all BFSs.

2.2.2 Node Measures

Node-based centrality [1] is defined in order to measure the importance of a node in the network. A node with high centrality score is usually considered more highly influential than other nodes in the network. The main principle to categorize the centrality measures is the type of random walk computation involved. In particular, the centrality measures can be grouped into two categories: radial and medial measures. Radial measures assess random walks that start or end from a given node. On the other hand, medial measures assess random walks that pass-through a given node. The radial measures are further separated into volume measures and length measures based on the type of random walks. Volume measures fix the length of walks and find the volume (or number) of the walks limited by the length. Length measures fix the volume of the target nodes and find the length of walks to reach the target volume. Some popular centrality measures based on these categories are defined here.

- ***Degree Centrality***

The first group of the centrality measures is that of the radial and volume-based measures. The simplest and most popular measure in this category is that of degree centrality [1]. The importance of a node is determined by the number of nodes adjacent to it. One way of interpreting the degree centrality is that it counts the number of paths of length 1 that starts from a node. A natural generalization from this perspective is the K − path centrality which is the number of paths of length at most k that start from a node. The larger the degree, the more important the node is. Only a small number of nodes have high degrees in many real-life networks. Let A be the adjacency matrix of a network, and d_i be the degree of node i. The degree centrality is given by,

$$C_D(v_i) = d_i = \sum_j A_{ij}$$

The normalized degree centrality is given by,

$$C'_D(v_i) = d_i/(n-1)$$

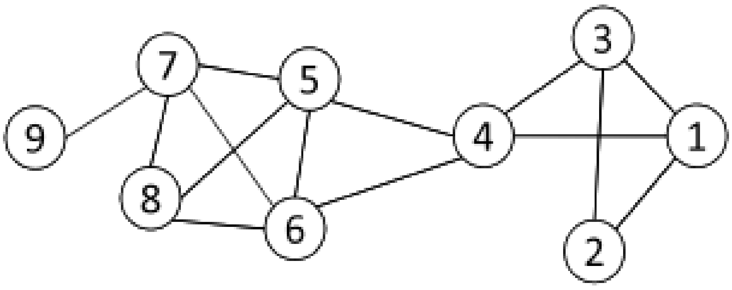


Fig-7

In fig-7 in the given graph for node 1, degree centrality is 3; Normalized degree centrality is 3/ (9-1) =3/8.

Freeman's general formula for centralization is used to determine how much variation is there in the centrality scores among the nodes.

$$C_D = \frac{\sum_{i=1}^{g}\left[C_D(n^*) - C_D(i)\right]}{[(N-1)(N-2)]}$$

, ***C_D (n^*)*** is the highest value in the network.

- ***Closeness Centrality***

The second group of the centrality measures is that of the radial and length-based measures. Unlike the volume-based measures, the length-based measures count the length of the walks. The most popular centrality measure in this group is the Freeman's closeness centrality [1]. It measures the centrality by computing the average of the shortest distances to all other nodes.

Closeness centrality measures how quickly a node can access more nodes in a network. In this measure importance is measured by how close a node is to other nodes.

Average distance is calculated using the following formula:

$$D_{avg}(v_i) = \frac{1}{n-1}\sum_{j\neq i}^{n} g(v_i, v_j)$$

Closeness Centrality is calculated using the following formula:

$$C_C(v_i) = \left[\frac{1}{n-1}\sum_{j\neq i}^{n} g(v_i, v_j)\right]^{-1} = \frac{n-1}{\sum_{j\neq i}^{n} g(v_i, v_j)}$$

***g* (*v_i*,*v_j*)** is the distance between node i and j.

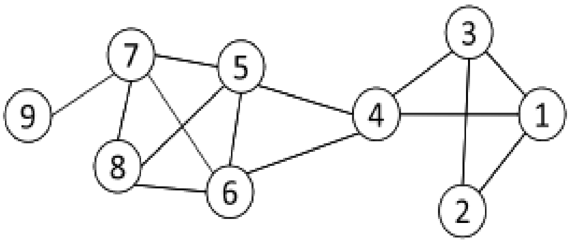


Fig-8

According to this graph in fig-8 the closeness centralities of nodes 3 & 4 are calculated below.

$$C_C(3) = \frac{9-1}{1+1+1+2+2+3+3+4} = 8/17 = 0.47,$$
$$C_C(4) = \frac{9-1}{1+2+1+1+1+2+2+3} = 8/13 = 0.62.$$

Therefore node 4 is more central than node 3.

- ***Betweenness Centrality***

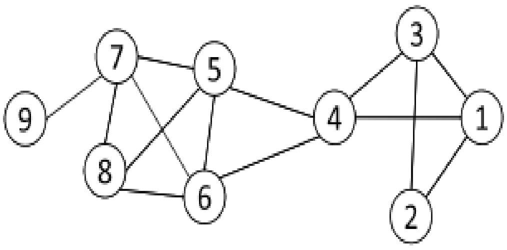
As is the case for edges of high betweenness, nodes of high betweenness occupy critical positions in the network structure, and are therefore able to play critical roles. This is often enabled by a large amount of flow, which is carried by nodes which occupy a position at the interface of tightly-knit groups. Such nodes are considered to have high betweenness. Another group of the centrality measures is that of medial measures. It is called ‘medial’ since all the walks passing through a node are considered. It measures how much a given node lies in the shortest paths of other nodes. It is related to nodes that span structural holes in a social network. Betweenness centrality [1] occurs when a node falls in a favoured position between two cliques in the network. Mathematically, Node betweenness centrality counts the number of shortest paths that pass one node. Nodes with high betweenness are important in communication and information diffusion.

Betweenness Centrality is calculated using the formula below.

$$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s<t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

σ_{st} : The number of shortest paths between s and t

$\sigma_{st}(v_i)$:The number of shortest paths between s and t that pass v_i .



	$\sigma_{st}(4)/\sigma_{st}$		
	$s=1$	$s=2$	$s=3$
$t=5$	1/1	2/2	1/1
$t=6$	1/1	2/2	1/1
$t=7$	2/2	4/4	2/2
$t=8$	2/2	4/4	2/2
$t=9$	2/2	4/4	2/2

Fig-9 (a) *Fig-9 (b)*

In fig-9(a) a graph is taken as an example to find the betweenness centrality of node 4, and fig-9(b) shows the necessary calculations.

$$C_B(4) = 15$$

- ***Eigen-vector Centrality***

Here importance of a node is determined by friends of it. If A has many important friends, one should be important as well.

$$C_E(v_i) \propto \sum_{v_j \in N_i} A_{ij} C_E(v_j)$$

$$\mathbf{x} \propto A\mathbf{x} \Longrightarrow A\mathbf{x} = \lambda\mathbf{x}.$$

The centrality corresponds to the top eigenvector of the adjacency matrix A.A variant of this eigenvector centrality is the PageRank score.

There are different considerations for modelling influence in social networks. In addition to edge and node strengths, the followings are additional analytical considerations relevant to social influence analysis [2]:

- *Multitopic*: Social influence will have different effects on different topics discussed in the social network. For example, assume two neighbours A, specialized in data mining and B, specialized in programming. A will have high influence on B when the topic is related to data mining while B will have higher influence on A when the topic is related to programming.
- *User actions*: Considering user actions and past behaviours while measuring influence.
- *Scalability*: The number of nodes in Online Social Networks increases rapidly. Therefore, there is a need to develop methods that scale well with large data sets.

Chapter 3

Present State of Art

In the process of problem formation our main motivation is to maximize the influence in a social network. Here the basic models of influence are discussed in details and a prototype is shown. Afterwards the community detection algorithm implemented in the experiment is discussed in details. Also, the later part of the chapter describes the 'compath' algorithm [8] which will give better results compared to other metrics.

3.1 Problem Statement

Given a directed and edge-weighted social graph $G = (V, E, p)$, a diffusion model m, and an integer $k \leq |V|$, find a set $S \subseteq V$, $|S| = k$, such that the expected influence spread $\sigma_m (S)$ is maximum.

Important terms regarding the problem statement are:

Active node: Influenced nodes are active node otherwise it is inactive

Seed set $\boldsymbol{S}$: Initial set of nodes selected to start the diffusion

Influence spread $\boldsymbol{\sigma (S)}$: Expected number of active nodes at the end of diffusion process, if set S is the initial active set.

The solution approach of the above problem may be conceptualized as a collection of the following three steps:

a. Finding disjoint communities within a given network. Here community detection is vital because the target audience is easily identified and the impact of influence can be maximized. In real life situation viral marketing or many other business models follow such process to separate the whole network into different communities.
b. After the communities are formed the next task is to find seed nodes set from every community depending on the size of the community using well known metrics and modified algorithm (compath).
c. Finally, the independent cascade model is used to spread the influence throughout the network.

In the following part the basic models and algorithms are discussed in details and the most efficient model is implemented in the experiment.

3.2 Models Related to Influence Maximization

The categorization of social influence models are as follows: 1) static models, 2) dynamic models, 3) diffusion models, and 4) models based on users' behaviours.

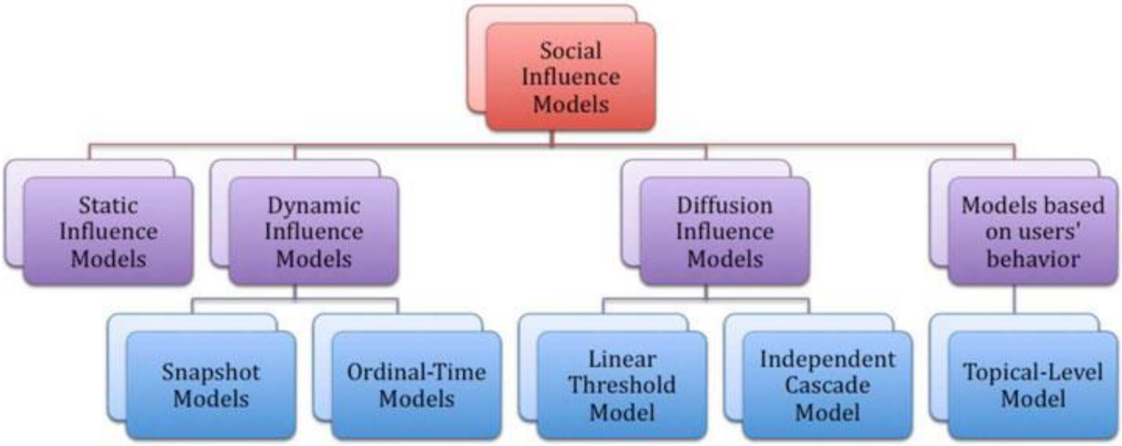


Fig-10

3.2.1 Static Influence Model

Static influence models [2] are independent of time and used to capture the most influential nodes presently. Therefore, the network size is fixed. One instance of this model is based on Bernoulli distribution. In these social influence models, a specific node u has a fixed probability to influence its inactive neighbour v. If it activates the neighbour r, then this is a successful attempt and otherwise, failure. Each attempt can be shown as a Bernoulli trial.

3.2.2 Dynamic Influence Model

In real life, influence changes over time and may not stay static. For example, users' opinions could change over time. When a user is influenced by its neighbours to join a community, she/he is initially excited to join that community, but over time that user might have less excitement to stay in the community. To represent dynamic influence models, there are two models of social influence. The first one is based on capturing a small set of "Snapshot" observations of the social network and the second one is based on detailed temporal dynamics. These two models can be represented as a function of the number k of neighbours who have adopted a new behaviour. The individual become k − exposed to the behaviour at specific time t if it is a nonadopter at time t but surrounded with exactly k neighbours who are all adopters at time t.

3.2.3 Diffusion Influence Model

These models are used when adopting behaviour depends on knowing the number of neighbours who adopted the same behaviour.

Directed graph G = (V, E)

Node $v \in V$ represents an individual

Edge $(u, v) \in E$ represents the influence relationship

Discrete time t: 0, 1, 2,

State of nodes: active or inactive

S_t: set of active nodes at time t

S_0, seed set, initial nodes selected to start the diffusion

- ***Linear Threshold Model***

Each edge (u, v) has an influence weight w (u, v):

if $(u, v) \notin E$, $w(u, v) = 0$

$\Sigma_u w(u, v) \leq 1$

Each node v selects a threshold $\theta_v \in [0,1]$ uniformly at random.

Initially seed nodes in $S_0$are activated.

At each step, node v checks if the weighted sum of its activated in-neighbours is greater than or equal to its threshold θ_v, if so, v is activated

At time stamp t, all nodes that were active in time $t - 1$ remain active, and any node v for which the total weight of its active neighbours is at least θ_v gets activated; where:

$$\Sigma_u w(u, v) \geq \theta_v$$

The thresholds θ_v represent the tendency of nodes to adopt the new behaviour when their neighbours do.

- ***Independent Cascade Model***

Each edge (u, v) has an influence probability p (u, v) and seed nodes in S_0 are activated

The process occurs in discrete steps as follows:

When node u becomes active for the first time in time step t, it provided with one chance to activate each of its currently inactive neighbour v;

In that case u is called contagious, which means that it has the ability to affect other nodes. Node u succeeds to influence its neighbour v with probability p (u, v) independent of past history.

If u succeeds, then v will become active in time step t + 1 but whether or not u succeeds, it cannot make any further attempts to activate v in future rounds. The same process continues until us communicate with all neighbours for influence attempts and there are no more contagious nodes.

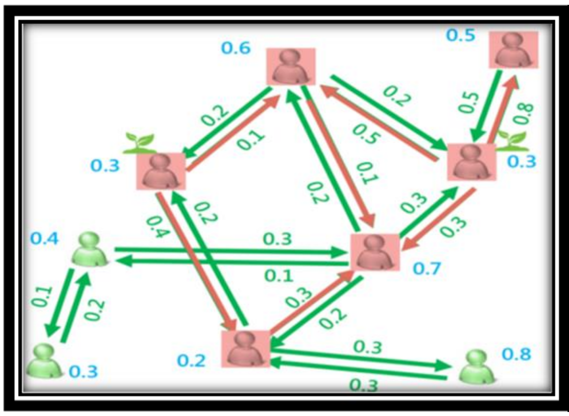


Fig-11(a) Linear Threshold Model

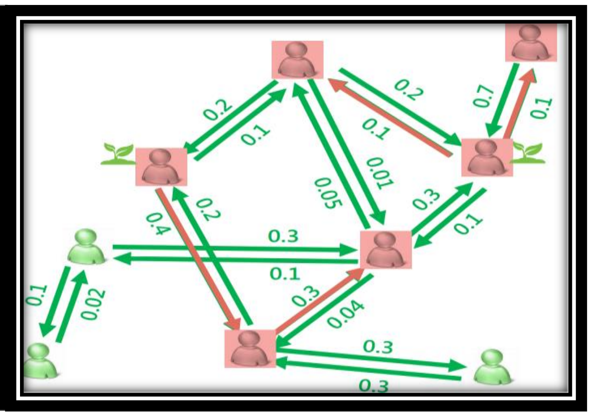


Fig-11 (b) Independent Cascade Model

3.2.4 Models of Influence based on Users' Behaviour

The models discussed above are based on the influence model where the influence probabilities are provided in advance as input. Other models proposed compute the probabilities through mining past users' behaviour.

Topic-based social influence is one of them. In these social networks, discussion topics are distributed across users. The problem is then to find topic-specific subnetworks, and topic-specific influence weights between members of the subnetworks.

3.3 Small Prototype to understand the process of influence spreading

- Let us suppose we have a graph where nodes represent students and the edges represents friendship.
- There are two ideas: A. To go out for movie (payoff/ willingness=20)
  B. To study in library (by default everyone's state) (payoff=10)

Which nodes we should select and infect them with the idea A initially such that it then influences others? We also make sure these seed nodes are the least in number making maximum impact.

- Consider the graph:

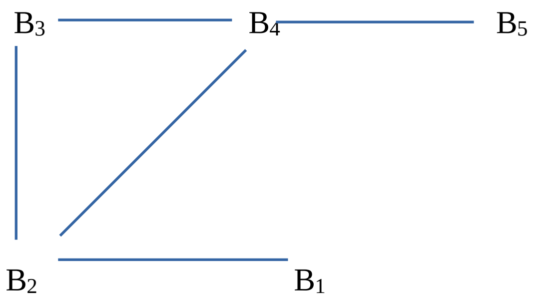


Fig-12 (a)

- Here all nodes have idea B. Now we infect B_2 and B_4 with idea A initially. We choose them due to their high degree centrality.

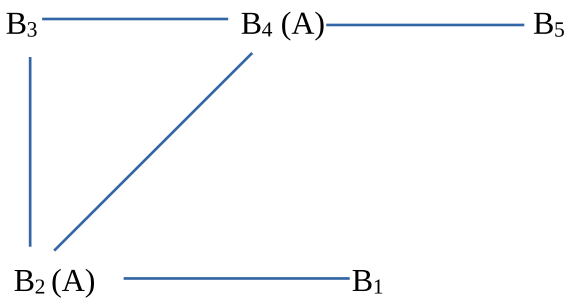


Fig-12 (b)

- For 1st iteration we check all the nodes of the graph if they accept the idea of A.
- B_1==> No. Of neighbours having idea A= 1=> 1* payoff=1*20=20.
- B_1==> No. Of neighbours having idea B= 0=> 0* payoff=0*10=0.
- Therefore B_1 accepts idea A as 20>=0.
- B_2==> No. Of neighbours having idea A= 2=> 2* payoff=2*20=40.
- B_2==> No. Of neighbours having idea B= 1=> 1* payoff=1*10=10.
- Therefore B_2 accepts idea A as 40>=10.
- B_3==> No. Of neighbours having idea A= 2=> 2* payoff=2*20=40.
- B_3==> No. Of neighbours having idea B= 0=> 0* payoff=0*10=0.
- Therefore B_3 accepts idea A as 40>=0.
- B_4==> No. Of neighbours having idea A= 2=> 2* payoff=2*20=40.

- B_4==> No. Of neighbours having idea B= 1=> 1* payoff=1*10=10.
- Therefore B_4 accepts idea A as 20>=0.
- B_5==> No. Of neighbours having idea A= 1=> 1* payoff=1*20=20.
- B_5==> No. Of neighbours having idea B= 0=> 0* payoff=0*10=0.
- Therefore B_5 accepts idea A as 20>=0.

Hence, we see the cascade is complete as all the nodes accepted the idea A.
Now we carry forward this method to apply it on the below described datasets and continue further to implement the algorithm to find the most influential seed node set spreading maximum influence in least time suitable for such influence models.

3.4 Algorithms used

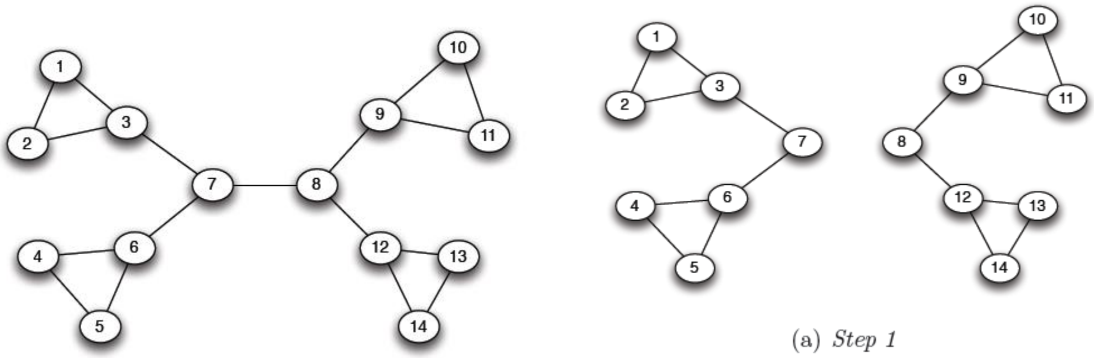
3.4.1 Girvan-Newman Community Detection Algorithm

It is a divisive hierarchical clustering algorithm based on edge betweenness [1]. The main steps of the algorithm are as follows –

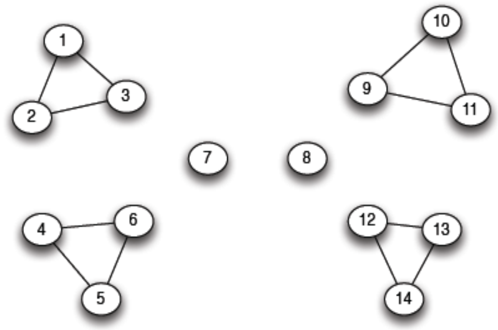
i. The betweenness of all existing edges in the network is calculated first.
ii. The edge with the highest betweenness is removed.
iii. The betweenness of all edges affected by the removal is recalculated.
iv. Steps ii and iii are repeated until no edges remain.

The end result of the Girvan–Newman algorithm is a dendrogram. As the Girvan–Newman algorithm runs, the dendrogram is produced from the top down (i.e. the network splits up into different communities with the successive removal of links). The leaves of the dendrogram are individual nodes. The connected components are communities.

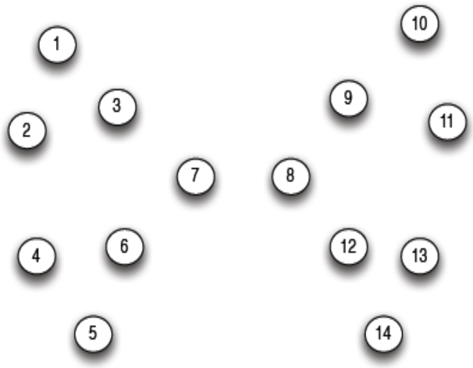
The below graph is initially completely connected but after edge removal in consequent steps1,2,3, the communities are found in hierarchical order.



(a) *Step 1*



(b) *Step 2*



(c) *Step 3*

Fig-13

3.4.2 Modified *Compath* Algorithm

The main steps of the algorithm [8] are summarized here.

First, community structures of the input graph are used to form a new network. Each node of the new network represents a community. Then the most central nodes of the original network are sorts based on the closeness centrality measure. Each node of the new network contains the nodes of its corresponding community. So, proportional to the size of the community, a number of nodes are selected from each community based on closeness centrality to from candidates set. Then the selected nodes are merged from the candidates set based on closeness centrality measures.

Step 1: Compute New_Network

In the first step of the algorithm, community structures of the input network are identified and then a reduced network is built as follows: first Girvan-Newman algorithm is used to detect communities of the input network then each community is considered as a node in a new graph G=(V,E). Where V is the set of nodes (communities) and E is the directed edges that exist between the graph nodes. So, a directed edge (A, B) shows that members of the community A are linked to members of B in the original network.

(a) (b)

. (a) *A sample input graph with two communities and* (b) *its corresponding reduced graph*

Fig-14 (a & b)

Step 2: ComputeSelected_Nodes

After construction of the communities' graph, the original graph nodes are sorted based on their closeness centrality measure. This measure shows the importance of each node in the input network. After choosing important communities, some representative nodes should be selected from each of them. Closeness centrality is a simple and efficient measure for this purpose according to our findings. Now, number of nodes that should be chosen from each community is an important work. A community that contains more nodes should play a more important role in the final result. The following formula is applied for this purpose: $\left\lceil \frac{N_i}{(Max-Min)} * \beta \right\rceil + \alpha$

In which Max is the number of nodes of the largest community, Min is the number of nodes of the smallest community and N_i is the number of nodes of the i^{th} community. The value of $\frac{N_i}{(Max-Min)}$ falls between zero and one, so it is multiplied by a constant β to amplify its effect. Also, α is a constant that ensures the selection of at least α nodes from each community. The appropriate value of α depends on the input network.

After that according to closeness centrality measurements and this formula the representative nodes from each community are selected and they are merged to form the seed set.

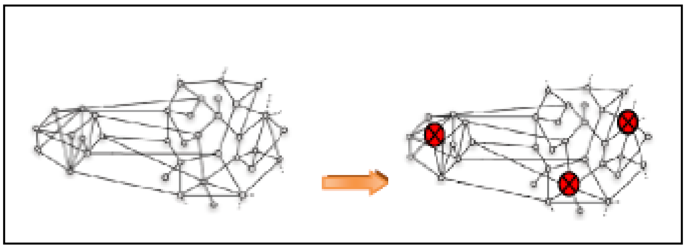


Fig-15 shows the selection of seed nodes from every community

Chapter 4

Experiments and Results

The further work includes the experimental work and studies related to influence maximization. Different metrics efficiency is compared later.

4.1 Assumptions

1. The network is static network with no incoming or leaving nodes in the network.
2. The influence spread mechanism is considered to be spreading only on the basis of neighbours being active or inactive.
3. The payoff is a measure which is assigned to every node and change according to the number of neighbours having accepted the idea or rejected the idea.
4. The nodes initially chosen as seed nodes remain active till the acceptance payoff becomes less than rejection payoff.
5. The algorithm helps to find the smallest set of seed nodes to maximize the influence via this process which may not be the case in general social network where other attributes play a role in having a positive or negative impact.
6. No sentiment analysis or other psychological behaviour of a node is taken into consideration. These has been summarized to the simple single metric of payoff.
7. Dynamic network or frequent topology change of general social network is not considered. Here it is a snapshot of the social network at a single time instant.

4.2 Datasets

Two real networks are used in our experiments:

I. The first dataset is a small static dataset of 100 nodes having non-overlapping communities having the edge list in a csv file with given weights and connectedness.

II. This static dataset consists of ‘circles’ (or ‘friends lists’) from Facebook. Facebook data was collected from survey participants using this Facebook app. The dataset includes node features (profiles), circles, and ego networks. It has highly overlapping communities.

Dataset statistics	
Nodes	4039
Edges	88234
Nodes in largest WCC	4039 (1.000)
Edges in largest WCC	88234 (1.000)
Nodes in largest SCC	4039 (1.000)
Edges in largest SCC	88234 (1.000)
Average clustering coefficient	0.6055
Number of triangles	1612010
Fraction of closed triangles	0.2647
Diameter (longest shortest path)	8
90-percentile effective diameter	4.7

Table-1

4.4 Experiments

4.4.1 Experiments with modified algorithm [compath] on Dataset-1

The dataset-1 is having no overlapping communities and there are a smaller number of nodes in the network which gives faster result when executed on limited capacity machine.
The outputs are as follows:
First the total communities are detected from the total network using Girvan Newman algorithm.

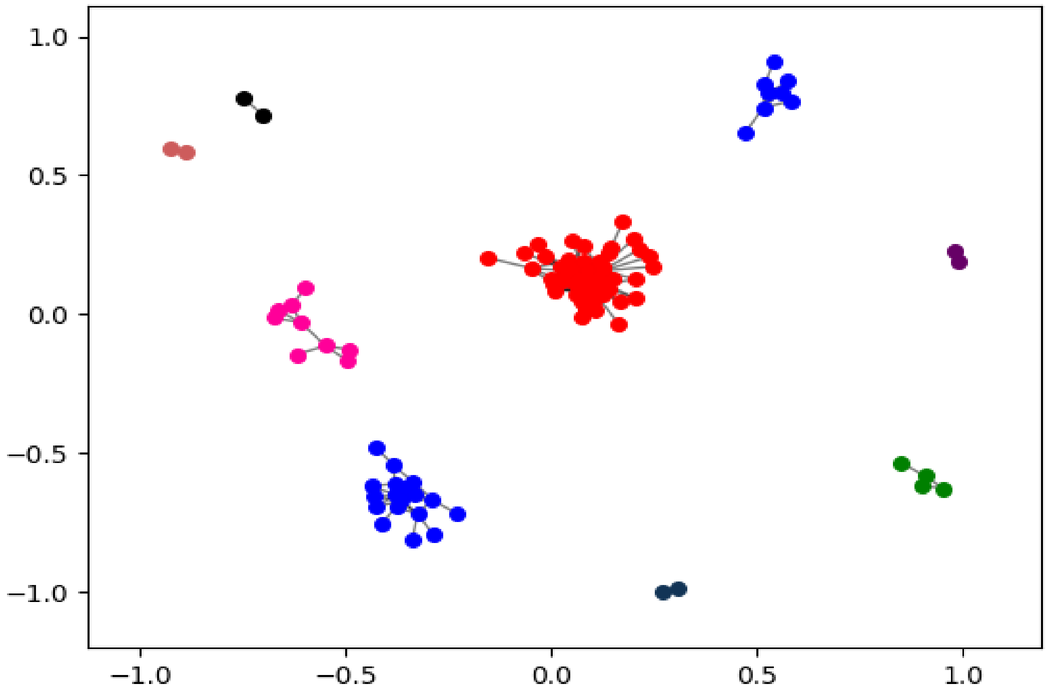


Fig-16: The graph shows the community formation till the modularity value increases.

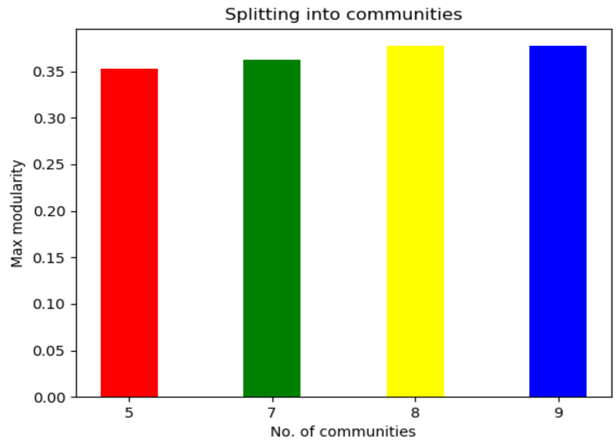


Fig-17: This graph shows that after 8 communities are formed the value is stagnant. So, iteration stops.

After community detection the size of seed node set from every community is calculated according to the size of the community itself.

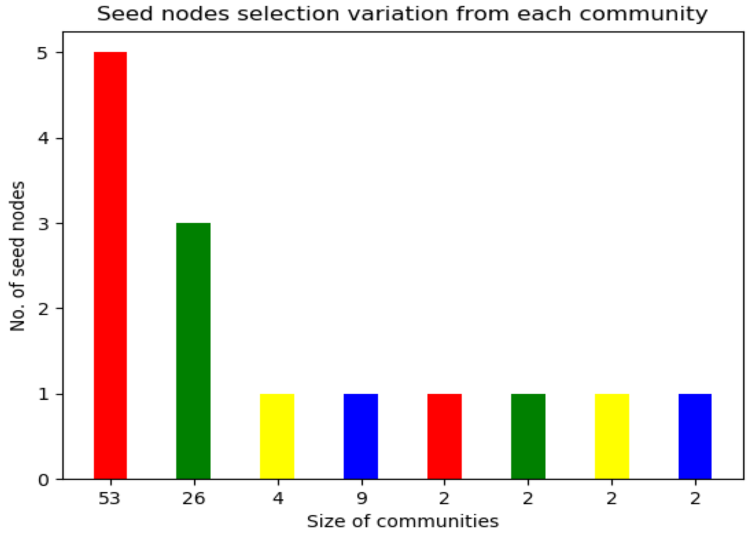


Fig-18

Then these 14 nodes are selected as the total seed nodes, on the basis of higher closeness centrality, which are to be made active initially to start influencing the total network.

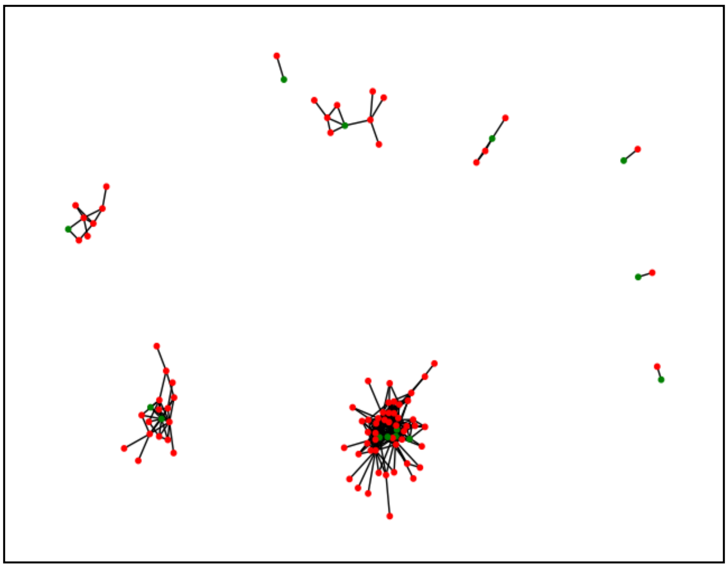


Fig-19: The initial network structure where 14 nodes out of 100 nodes are active.

After the selection is done the cascading model starts and the influence spreads throughout the network as shown in fig-20.

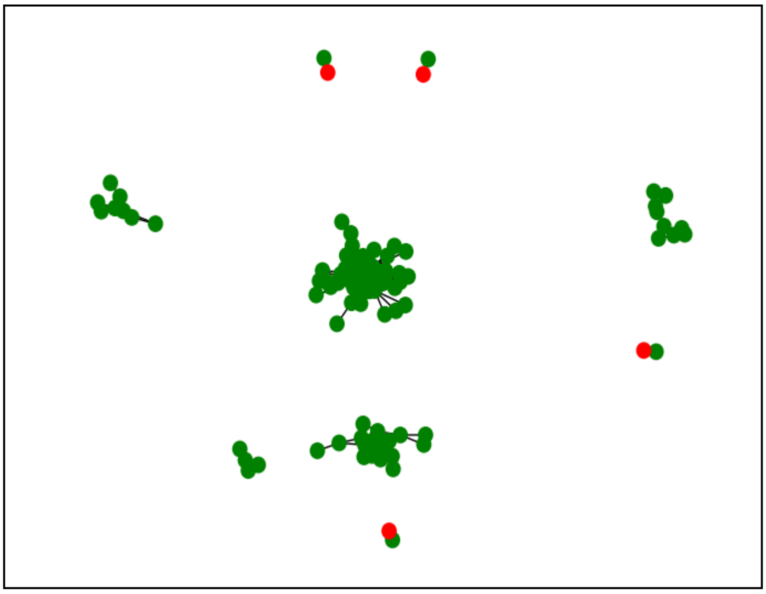


Fig-20: The final structure of the network where 96 nodes are active.

```
File  Edit  View  Search  Terminal  Help
Max modularity (Q): 0.361924
No of communities in decomposed G: 8
Modularity of decomposed G: 0.377108
Max modularity (Q): 0.377108
No of communities in decomposed G: 9
Modularity of decomposed G: 0.373531
Max modularity (Q): 0.377108
A community of length 53 :  {'Bharat7140', '_GuruGhantal_', 'TavisShahnazar', 'YourGirlNeha', 'AnkurYa20727560', 'HakimSinghRagh1', 'rahulyada
v0421', 'IAmShereAalam', 'MdKhan7861', 'Fairystylesmoon', 'pal_sahab74', 'Drraj94397198', 'Sajjad29101278', 'GunjaSV', 'aftabnazim786', 'Firoz
al60902021', 'talha4lee', 'ArshadChishti_', 'Maharani451', 'ERRKVAID3', 'ahanafatima786', 'azamnawaaz', 'Dilsedesh', 'YashodanandanS', 'kamran
YC', 'WDhongre', 'WasiuddinSiddi1', 'SSP_Twitt', 'alishaikh1737', 'Mushtaque51', 'IndiaPolls4', 'AaliyaAkhter', 'vikascongress', 'Rturki0786',
 'AzmiAxpress', 'akhterzaidi', 'TiwariSN5', 'hisamuddink51', 'kawaljeetyadav7', 'SUBASH50693', 'AbidDairdwal', '_jr_alam', 'AbrarNoori8', 'pra
fupriti', 'devendranathyad', 'Sana__Siddique', 'MOHAMMED___SRH', 'Rahis29411164', 'CongressALL', 'AnilKumarSaga11', 'Kadirsolanki2', 'MahapatP
ravat', 'Sayed23219671'}
A community of length 26 :  {'ReshmaLalwani6', 'RajTiwa56444261', 'ajaysdl1993', 'Aoaoao098', 'PawanSharma2004', 'BJPGMDassPondy', 'AnupamMish
ra8', 'Prerna_Thakur25', 'Mi_Ganu', 'Dipu11977Kumar', 'chetans88666', 'vera_anz', 'satishshuklaa', '56perumal', 'sachinitp', 'singhsantosh98',
 'HariOmDass7', 'RahulDubeyBABA', 'DharRenuka', 'SarlaSungroya', 'prav2410', 'AryanAArya9', 'vsarrulmozhy', 'V5336Nand', 'R_parab', 'singh_76'
}
A community of length 4 :  {'TechguruUp2', 'adityak58996261', 'AkhileshFan', 'akhileshcares'}
A community of length 9 :  {'Preeti354', 'Khushbu08043662', 'AnitaDasi12', 'ramesh00302', 'PreetiDasi9', 'BirjuKu04', 'BRamdevi', 'SarwanK3171
0137', 'U8yQTbfUkcPXk9m'}
A community of length 2 :  {'sachinsingh2305', 'bharati_mrinal'}
A community of length 2 :  {'hadhai_yaar', 'scrrwer'}
A community of length 2 :  {'DasiTwinkle', 'SmritiRani13'}
A community of length 2 :  {'_akhil_d', 'IamArvnd'}
number of nodes to be taken from the community are: 5
number of nodes to be taken from the community are: 3
number of nodes to be taken from the community are: 1
number of nodes to be taken from the community are: 1
number of nodes to be taken from the community are: 1
number of nodes to be taken from the community are: 1
number of nodes to be taken from the community are: 1
number of nodes to be taken from the community are: 1
till now: ['Sana__Siddique', 'hisamuddink51', 'AaliyaAkhter', 'ERRKVAID3', 'devendranathyad', '56perumal', 'Prerna_Thakur25', 'ReshmaLalwani6'
, 'akhileshcares', 'Preeti354', 'bharati_mrinal', 'scrrwer', 'DasiTwinkle', '_akhil_d']
Total active nodes:  96 Total inactive nodes:  4
souvik@souvik-Inspiron-5559:/media/souvik/New Volume/sem8/project/method/codes$
```

Fig-21: A short snapshot of the communities formed and the influence spread throughout the network.

- For 2 iterations the resulting active node count out of the total 100 nodes:

Centrality used	14 seed nodes
Degree	70
Betweenness	80
Closeness	87
Eigen-vector	62
Compath	96

Table-2

Now to compare the results of different metrics used earlier with the modified method 'compath' a graph is plotted to show the significant increase in number of active nodes in the network in fig-22.

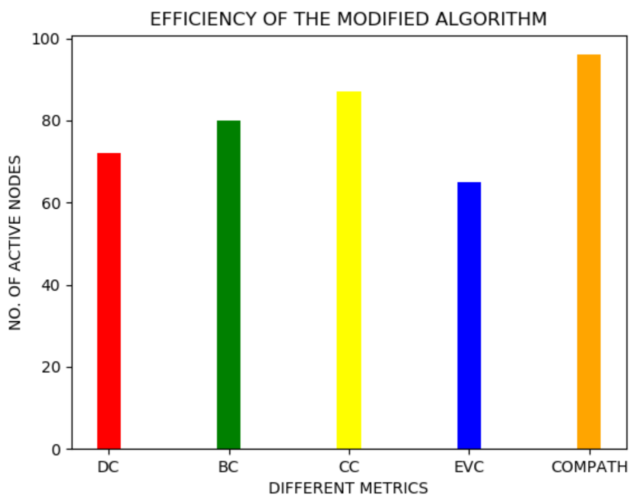


Fig-22

Hence from the results obtained it can be concluded that the compath algorithm it a much better algorithm to find the minimum seed nodes to spread maximum influence in the network. The number of active nodes is the highest for the 'compath' algorithm.

4.4.1 Experiments with modified algorithm [compath] on Dataset-2

For the Facebook dataset we have 4000+ nodes where only the edge list is given. Applying the modified algorithm of 'compath' to maximize the influence within the network more effectively. But the outcome is not significant as there are many overlapping communities. So after long execution time the following output is obtained.

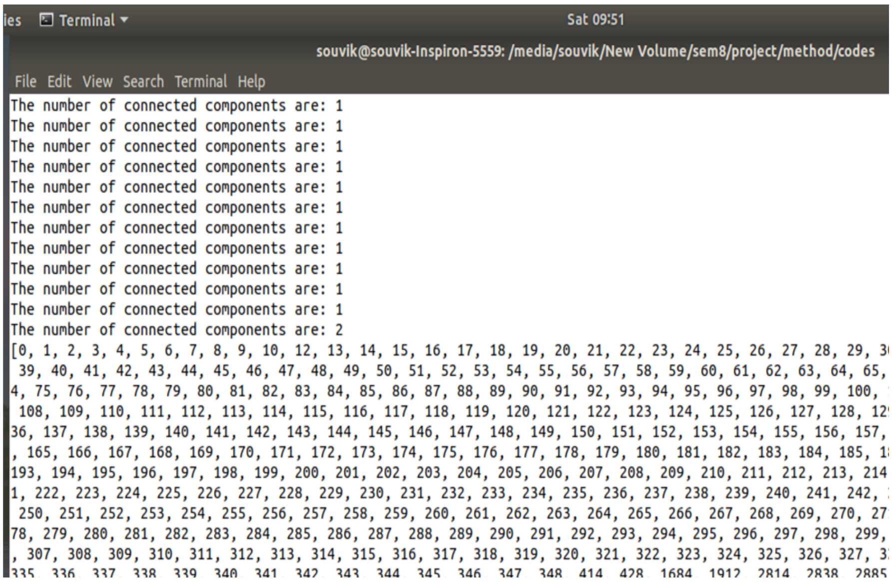


Fig- 23: Infinite iterations continue in search for communities

In this output it can be concluded that after many iterations the node having highest betweenness centrality gets separated into a different community. But all other nodes are so densely connected within the network that no more communities can be found. So due to this reason the 'compath' algorithm cannot work properly on huge network like this having overlapping communities.

So, the basic metrics are used in the dataset-2 and they are compared among themselves to check their efficiency. The influenced efficiency is tuned by increasing and fine tuning the total number of iterations or the random selection of size of seed node sets.
Initially 20 nodes are selected from the graph having the highest metric values, then we give the idea A to them so that they influence the other nodes to the maximum. Here as the graph size is huge the iterations may be huge to complete the cascade. So, for our purpose we restrict it to 100 iterations only. We select 20 influential nodes using four metrics and run the code four times and then compare them to find out which metric rightly predict the most influential nodes in this graph.
The initial phase and final phases of the network showing the active nodes are shown in the next pages in fig- 24-31.

- ***Degree Centrality***

Here we take those 20 nodes of the graph which have the highest degree centrality.

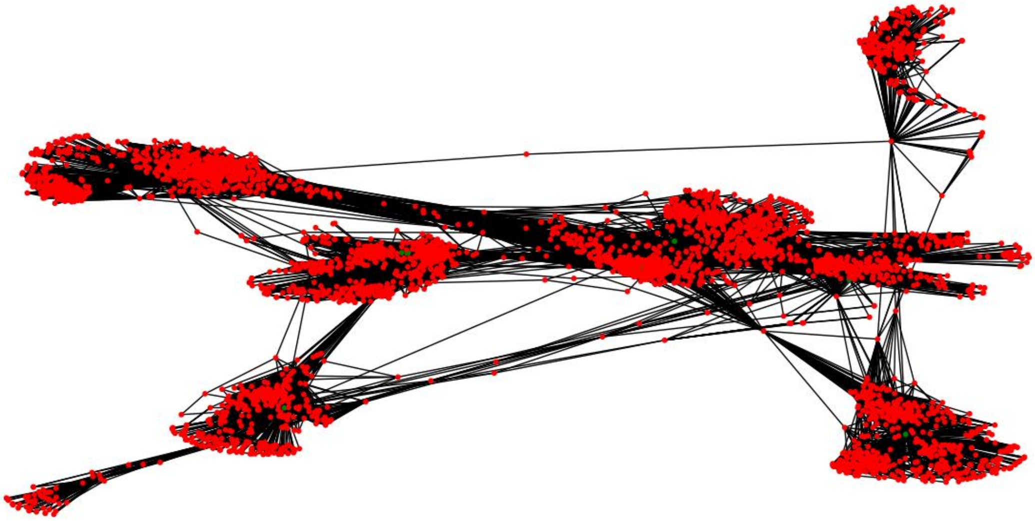


Fig-24: The initial status of the network having only 20 nodes active.

After 100 iterations:
Active nodes: 527

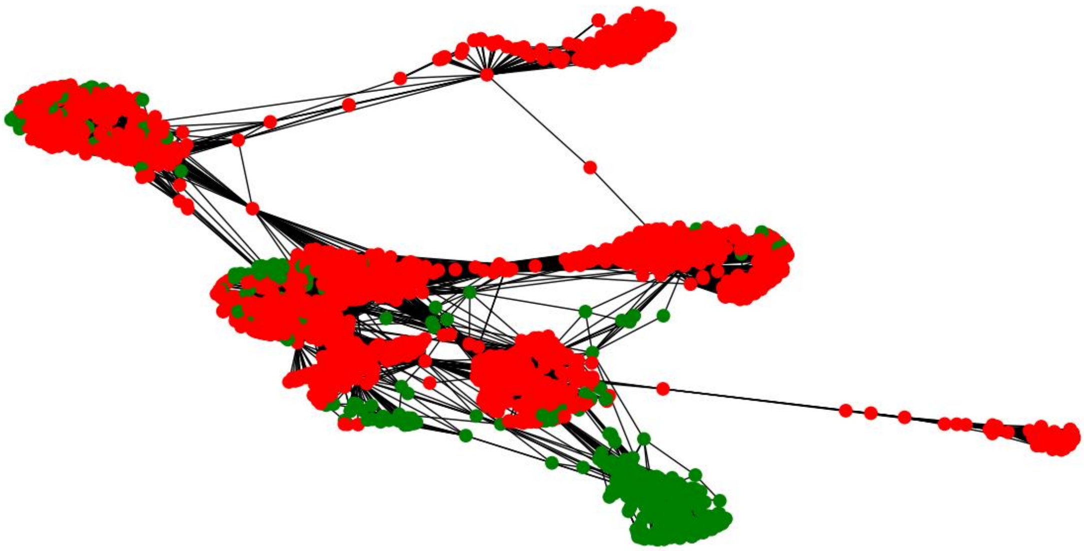


Fig-25: The final state of the network after activation of 527 nodes.

- ***Betweenness Centrality:***

Here we take those 20 nodes of the graph which have the highest betweenness centrality.

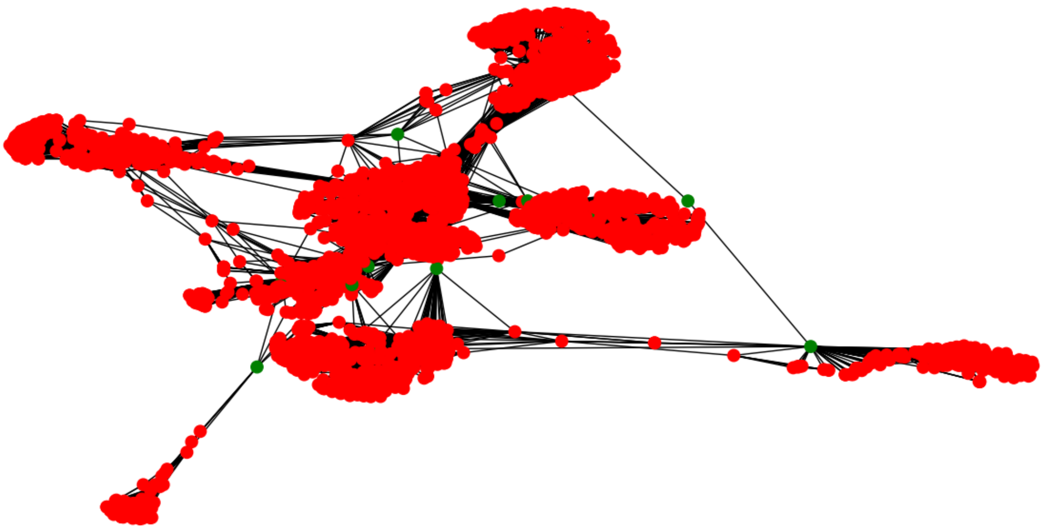


Fig-26: The initial status of the network having only 20 nodes active.

After 100 iterations:
Active nodes: 685

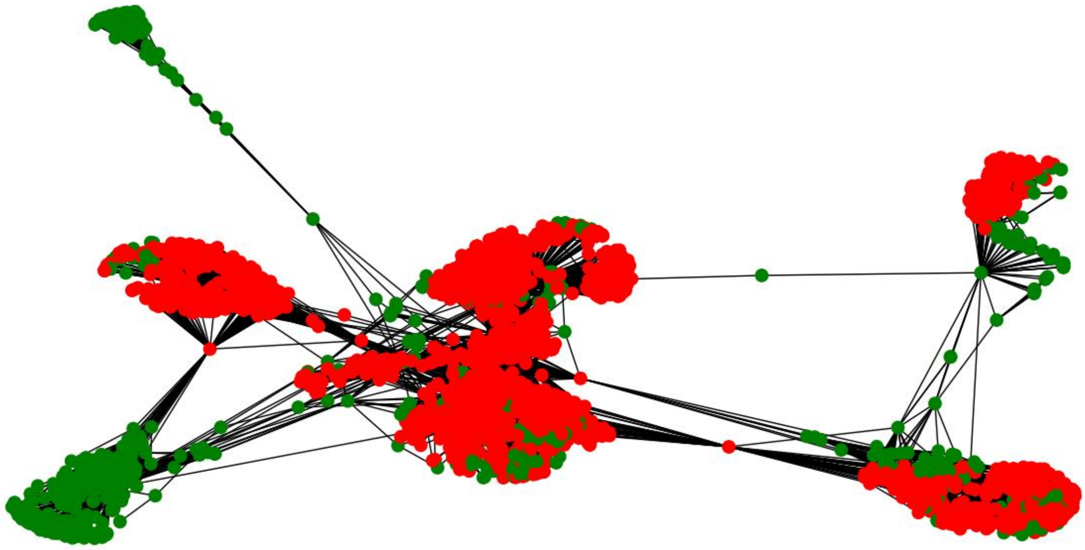


Fig-27: The final state of the network after activation of 685 nodes.

- ***Closeness Centrality*:**

Here we take those 20 nodes of the graph which have the highest closeness centrality.

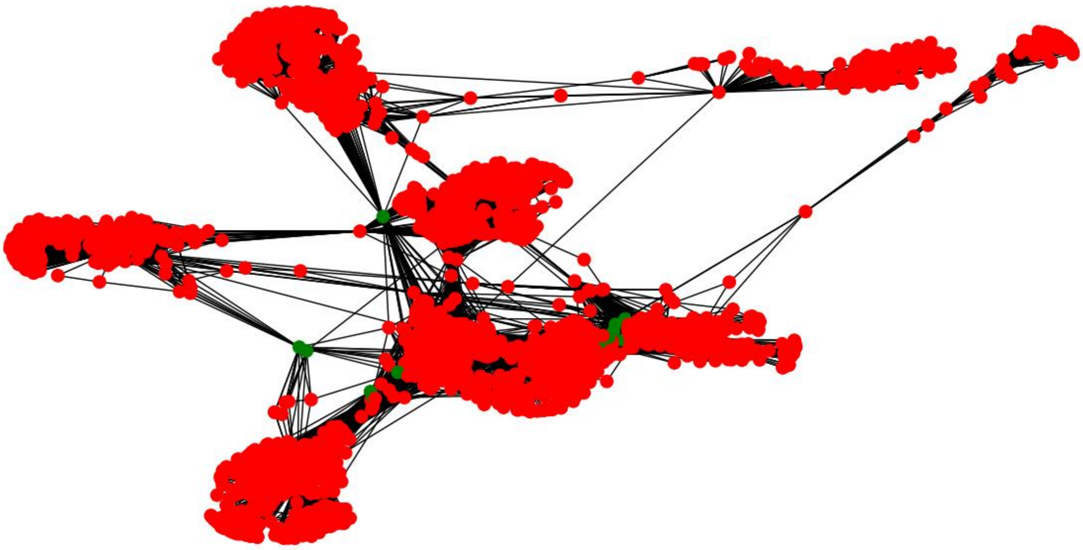


Fig-28: The initial status of the network having only 20 nodes active.

After 100 iterations:
Active nodes:1177

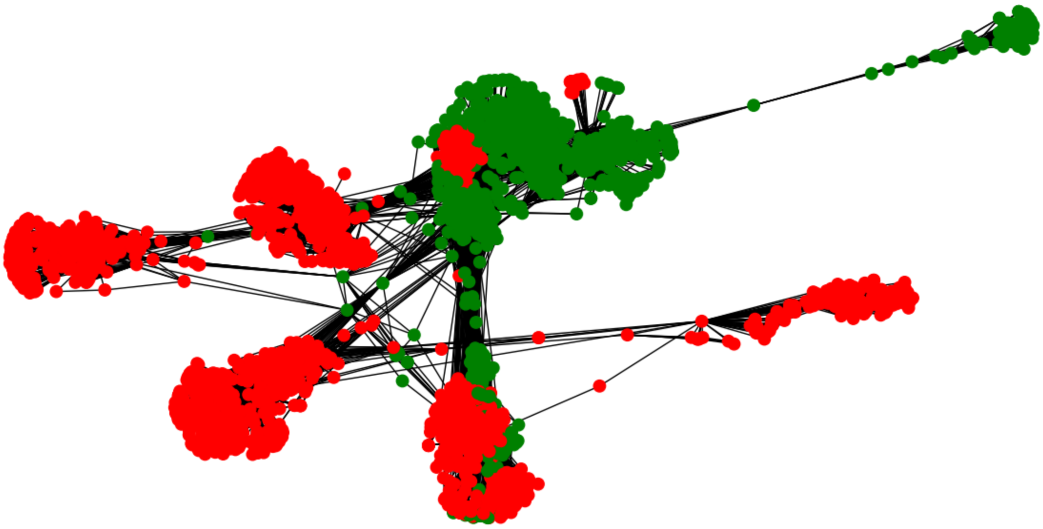


Fig-29: The final state of the network after activation of 1177 nodes.

- ***Eigen Vector Centrality*:**

Here we take those 20 nodes of the graph which have the highest Eigen vector centrality.

Fig-30: The initial status of the network having only 20 nodes active.

After 100 iterations:
Active nodes: 22

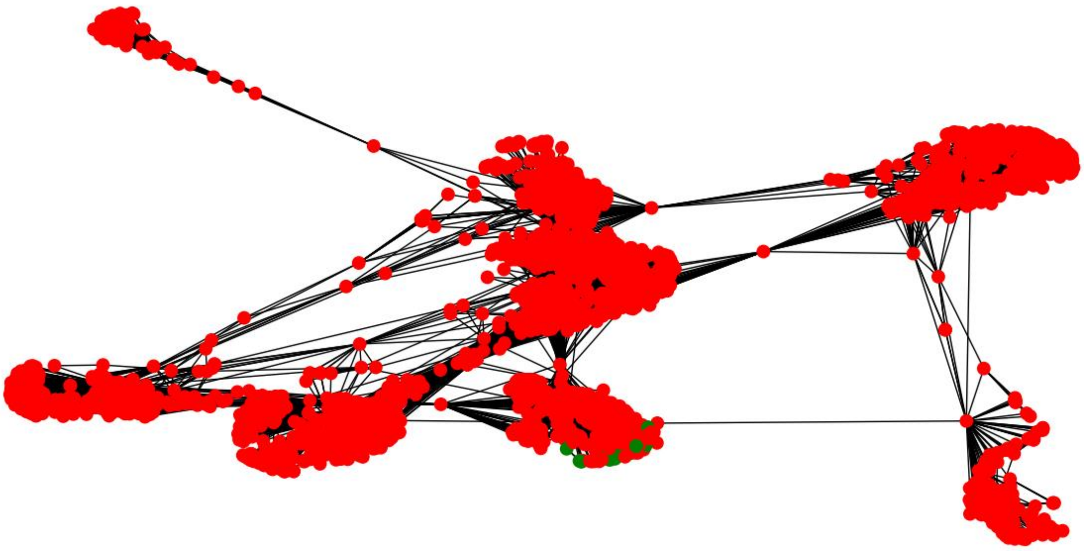


Fig-31: The final state of the network after activation of 22 nodes.

Now we compare all four metrics by plotting a graph of the total active nodes after 100 iterations in fig- 32:

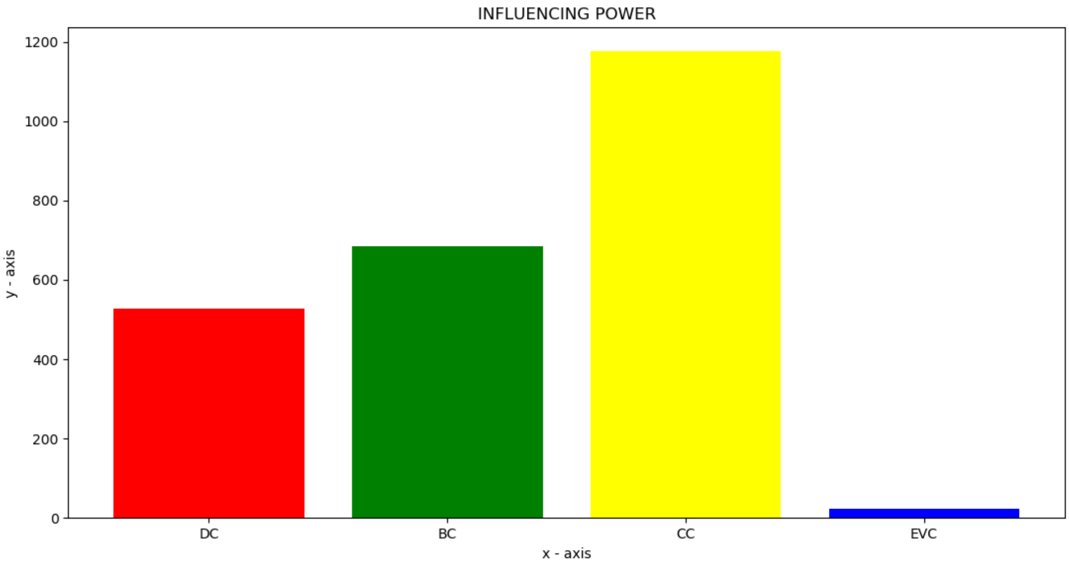


Fig-32

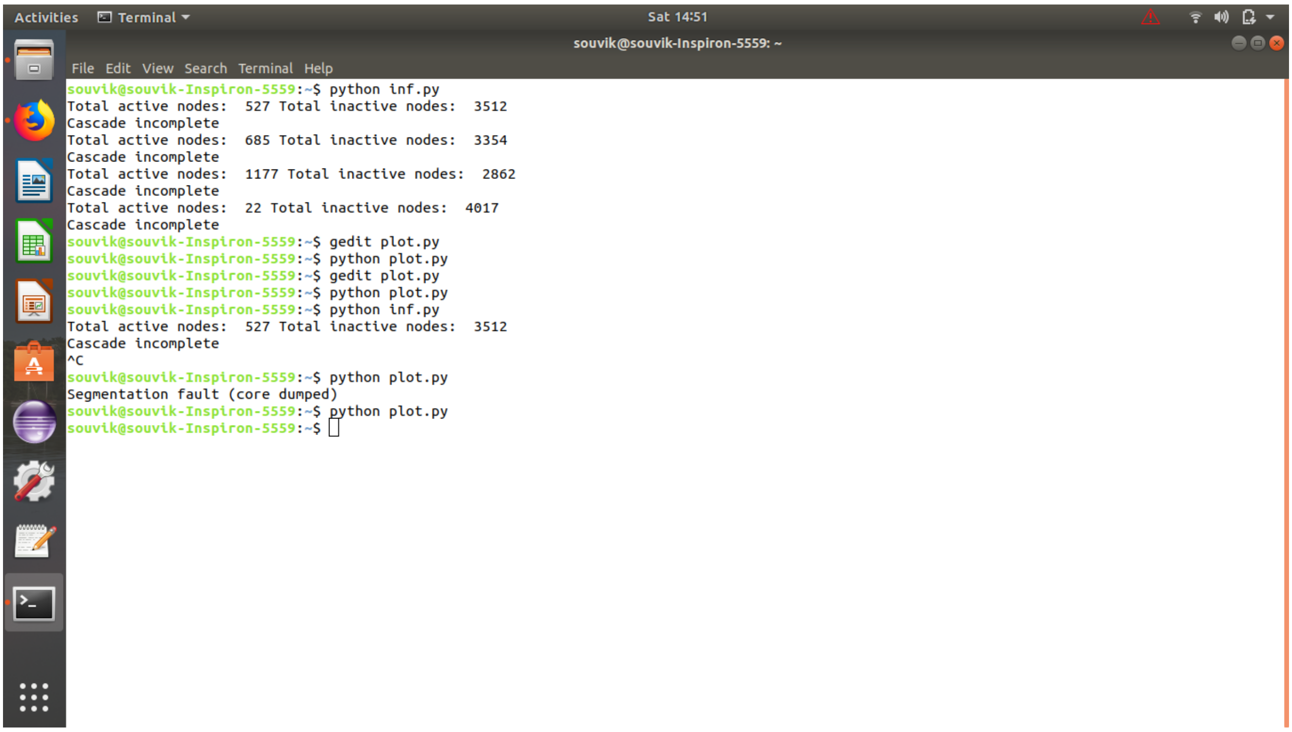


Fig-33: The completion of cascades for each metric after 100 iterations

Hence, we conclude that the closeness centrality is more important metric to predict the influential nodes in the graph. Thus, closeness centrality is considered for further experiments

- Here we can vary the two most important parameters of the algorithm for improving the count of active nodes in the network. These are as follows:

1. To increase the number of selected seed nodes.
2. To increase the number of iterations and persisting the continuous influence spread.

This method is applied to the same network with changing parameter values and noticing improvement in the model. We stop if the result starts degrading.
The table shows the variations in the increase and decrease of the active nodes in the network for certain parametric values:

- For 50 iterations the resulting active node count out of the total 4000+ nodes:

Centrality used	20 seed nodes	30 seed nodes	40 seed nodes
Degree	527	527	527
Betweenness	604	624	598
Closeness	872	1052	1052
Eigen-vector	22	452	322

Table- 3

- For 100 iterations the resulting active node count out of the total 4000+ nodes:

Centrality used	20 seed nodes	30 seed nodes	40 seed nodes
Degree	527	527	527
Betweenness	685	704	685
Closeness	1177	1538	1524
Eigen-vector	22	756	659

Table- 4

Result: Hence it is noted that the best result is obtained if we take 30 seed nodes and iterate the influence loop for 100 times. There may be some improvement for other values, but due to limited computation capacity we accept this result.

Chapter 5

Conclusion

Social influence maximization models aim to identify the smallest number of influential individuals (seed nodes) that can maximize the diffusion of information or behaviours through a social network. The influence maximization problem can be optimized further with some efficient greedy approach. In this project, an efficient algorithm, named ComPath, was used on the cascade model to find top-k most influential people in social networks. Our experiments showed the algorithm provides a good balance between execution time and efficiency. The algorithm detects the communities of the input network at first and investigates a limited number of them to reduce the execution time. Also, ComPath is able to suggest a limited number of communities to be investigated from the input network. This is very useful in our real world because it enables us to choose a few numbers of communities as our target and reduce the expenses for different applications such as product advertisement, etc. The method implemented in the project is not effective for highly overlapping networks which is a concern for future improvement. There are many possibilities in future to implement more sophisticated models on dynamic networks in real life. Perfect influence maximization model may be difficult to develop as the networks in real life is dynamic, and the influential probability changes with the change in ones thought process. Also, the monotonicity of a node is very hard to assure in real life scenario. Improved version of the algorithm can be obtained with more efficient greedy algorithms.

References

[1] KannaAlFalahi, Yacine Atif, Ajith Abraham, “Models of Influence in Online Social Networks”, International Journal Of Intelligent Systems, Vol. 29, pg- 161–183 (2014), Faculty of Information Technology, UAE University, Al Ain, United Arab Emirates, VSB-Technical University of Ostrava, Ostrava, Czech Republic.
[2] Jaewon Yang, EE Department, Stanford University, Jure Leskovec, “Modelling Information Diffusion in Implicit Networks “, CS Department, Stanford University. (2010)
[3] Sinan Aral and Paramveer S. Dhillon, “Social influence maximization under empiricalinfluence models” (2018), https://doi.org/10.1038/s41562-018-0346-z
[4] Ding-Zhu Du, “Influence Maximization”, Issue No. 01 - Feb. (2017 vol. 25), University of Texas at Dallas. Pp- 112-125 http://doi.ieeecomputersociety.org/10.1109/TNET.2016.2563397
[5] Chen, W., Wang, Y., & Yang, S. (2009). Efficient influence maximization in social networks. Paper presented at the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France.
[6] Goyal, A., Lu, W., & Lakshmanan, L. V. S. (2011b). SIMPATH: An efficient algorithm for influence maximization under the Linear Threshold Model. Paper presented at the the 2011 IEEE 11th International Conference on Data Mining.
[7] Wang, Y., Cong, G., Song, G., &Xie, K. (2010). Community-based greedy algorithm for mining top-Kinfluential nodes in mobile social networks. Paper presented at the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA.
[8] Rahimkhani, K., Aleahmad, A., Rahgozar, M., Moeini, A., A Fast Algorithm for FindingMost Influential People Based on the Linear Threshold Model, Expert Systems with Applications (2014), doi: http://dx.doi.org/10.1016/j.eswa.2014.09.037