Project Report on

**Flixster Network Analysis for Marketing**

by

**Nikita Chhajed**

**Sneha Tiku**

**Aditya Prithvinath**

**Abhishek Singh**



Introduction:

Flixster is a social networking site for movie fans. Users can create their own profiles, invite friends, rate movies and actors, and post movie reviews as well. From the site, people can also get information about movies, read user-generated movie reviews and ratings, converse with other users, get movie showtimes, view popular celebrity photos, read the latest movie news, and view video clips from popular movies and TV shows. With so many features, Flixster seems to have created a comprehensive movie social networking site, which may in turn explain its rapid growth in popularity. It provides user-item rating and user-user friendship networks.

What is social network Analysis? Social network analysis [SNA] is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships.

To analyze the different aspects of a social network, we must analyze the different connections in a network, the distributions of the nodes and edges, segmentation of human based on the relationships they share with each other and how network spreads.

Individual’s network grows as more people are accessing the social networking sites. Presenting interactions between the users, allows discovering important users as well as observing how influential they are. There are millions of social network users in the market, with thousands of connections. Hence, there is a need to organize these networks. One way to organize the networks, is to segment the users into different clusters based on their preferences, similar interests and characteristics. Users often take into consideration the recommendations and opinions of trusted people. Social network analysis helps in devising algorithms that can help in conducting successful advertising campaigns and predicting trends of popularity. Several models of influence have been proposed by social media, which will guarantee the maximal spread over the whole network.

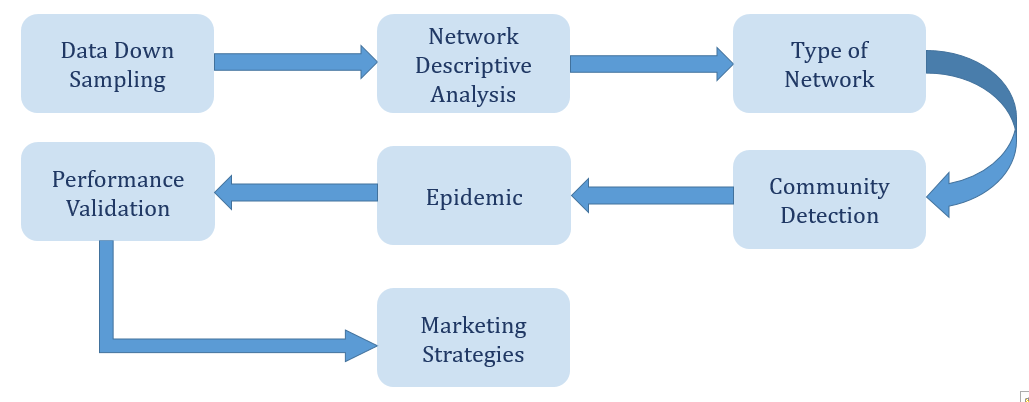
Statement of Purpose

In this project, we mainly focus on the following things:

1. To analyze the structure of the network and perform a detailed descriptive analysis to understand the connections
2. To apply different community detection algorithms to detect communities based on user preferences and movie ratings
3. To run an epidemic model on the network communities, to find the optimal advertising campaign duration for effective viral marketing

Approach for Analysis

Below is the detailed approach that we are going to follow for our analysis:



Network Structure:

Data source: <http://socialcomputing.asu.edu/datasets/Flixster>

We have a dataset that consists of two files.

1. Nodes.csv:

* This file contains 2,523,386 nodes
* It is the dictionary of all users who use Flixster

1. Edges.csv:

* This file contains 9,197,338 edges
* The edges indicate friendships among users

We used Gephi to visualize this network on Fruchterman reingold layout. Some of the observations of this network are as follows:

* Network Diameter = 19

This implies most distantly connected friends are separated by at least 19 friends

* Degree

User Id-21372 has a degree of 65. This implies that this user has the most number of users connected to him

* Average Clustering co-efficient

If the User 1 is connected to User 2 and User 2 is connected to User 3, then there is a 0.023 probability that User 1 is connected to User 3. This phenomenon is called transitivity

Below is the gephi representation of the network. The first graph is the full representation, while the second one is the zoomed version of the first graph, for better insights.

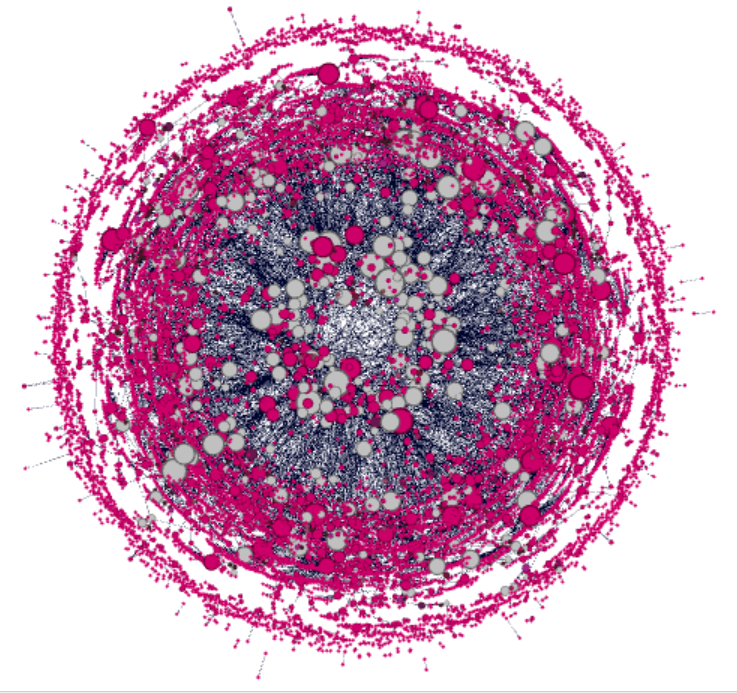


Fig 1 – Flixster Network - full

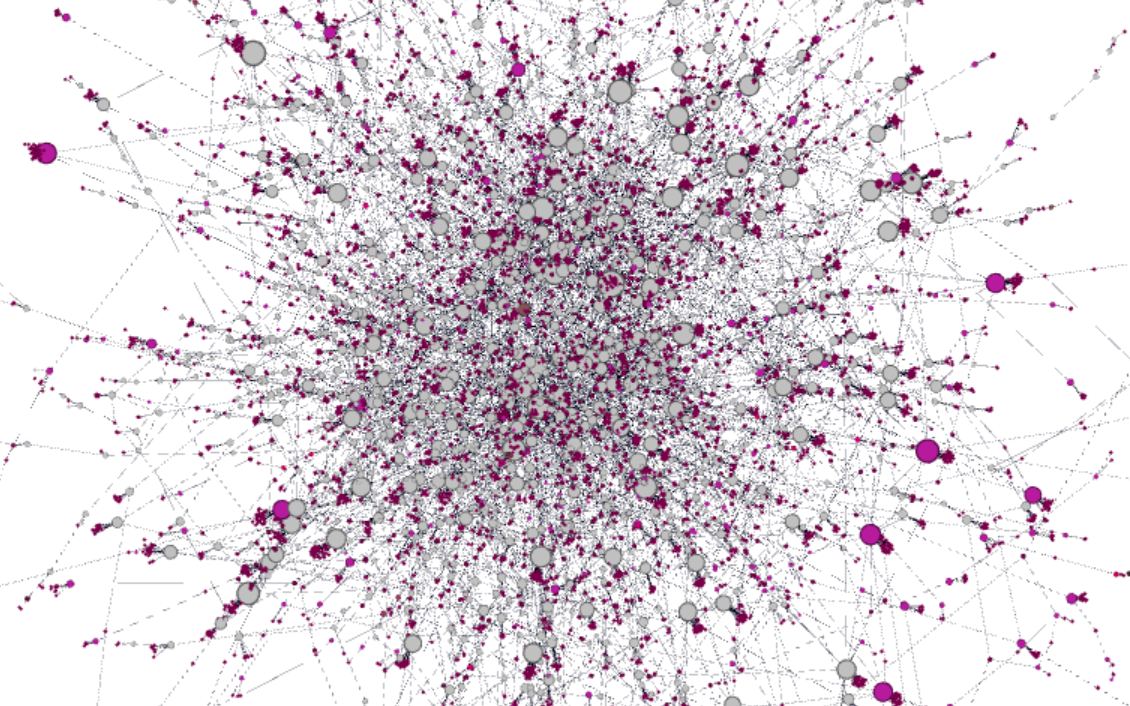


Fig 2 – Flixster Network - Detailed

Down-sampling:

This network is very huge to analyze and hence we decided to down sample the data and further do the analysis on the down sampled data.

The down-sampled data now has 1000 nodes and 4822 edges. Below is the network, created from the down sampled data. The size of the nodes is proportional to the degree of the nodes.

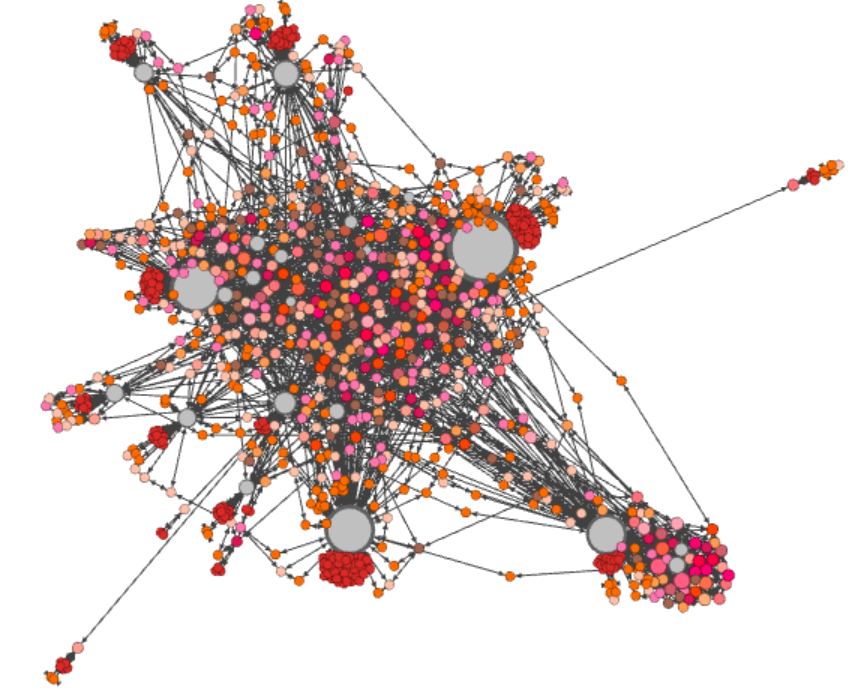


Fig 3: Flixster network – Down-sampled data

Network Metrics

Now we will carry out a detailed analysis of the Network metrics for this dataset.

* Network Diameter 🡪4

Diameter is the average distance between the pair of vertices. Graphs are conducted at random, subject to the degree sequence having the desired distribution and then the diameter is analyzed. In our graph, the diameter of the graph is 4.

* Average Path Length 🡪3.389

It is number of steps along the shortest paths for all possible pairs of network nodes. It is a measure of the efficiency of information or mass transport on a network.

* Average Degree Centrality 🡪4.826

This measure denotes degree centrality measures for all the nodes in the graph.

Centrality indicates how important a node is, i.e. it signifies the most influential partner in this Flixster network. The node with the highest degree is the most important partner with the largest number of ties in the network.

* Average Betweenness 🡪 1193.149

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes.

In this graph, it is a measure for quantifying the control of a user on the communication between other users in a social network. It measures how important one user is in influencing other users, to watch the same movies and rate them accordingly

* Transitivity🡪0.383

The Clustering coefficient captures the network’s local density at that node. The more densely interconnected its neighborhood is, the closer to 1 is its local clustering coefficient.

It is the estimate of the probability that each pair of its neighbors is connected.

* Assortativity 🡪-0.237

Assortativity is a preference for a network's nodes to attach to others that are similar in some way. Though the specific measure of similarity may vary, we can examine assortativity in terms of a node's degree.

Positive assortativity indicates a correlation between nodes of similar degree, while negative indicates a correlation between nodes of different degree.

Type of the Network

Community Detection

Entities of similar interests tend to be closer to each other. This is the primary reason for clusters formation. In real life, individuals studying at UIC will form a community based on the choice of courses they take, their interest in a major etc. Similar approach cascades to networks too. Community discovery in a social network has many applications. These include expertise finding and neighborhood query, and behavioral prediction. The structure of a community, which accounts for inherent dependencies between individuals in a social network, can help us understand the behavioral dynamics of individuals. Through characterizing multidimensional interpersonal relationships and an individual’s interests, community analysis can provide a quantitative summary of key factors related to word-of-mouth communications: tie strength, homophile, and source credibility. We can use community analysis to organize and to track content in online social media. There are significant opportunities for businesses: in addition to understanding user behavior and comments for better product design, businesses can take advantage of sentiment analysis of comments to proactively address negative commentary.

From the figure, we can see that there are few groups of closely connected nodes. We call these groups as clusters/communities. Not every network has such communities easily detectable by human eye. Also, a huge cluttered community might contain more than one nuclear communities inside it. We can make use of community detection algorithms which simplifies our task of finding communities in networks.

Some of the widely-used algorithms are

* Fast Greedy
* WalkTrap
* Springlass

Each of these algorithms have their own pros and cons. We have used the Walktrap Community detection algorithm to detect clusters in the Flixster Network. This algorithm works on the intuition that random walkers tend to get trapped into densely connected parts. The general idea is that if you perform random walks on the graph, then the walks are more likely to stay within the same community because there are only a few edges that lead outside a given community. Walktrap runs short random walks of 3-4-5 steps (depending on one of its parameters) and uses the results of these random walks to merge separate communities in a bottom-up manner.

Below is the output of the Walktrap algorithm on our dataset:

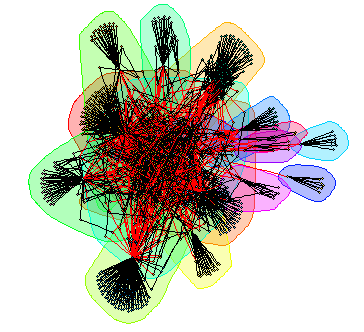


Fig 4: Walktrap Algorithm deployed in R

The above image clearly shows different clusters resulted from executing the algorithm on our network data. The clusters are formed based on the random walks the algorithm makes from each node. With a varied parameter settings in Gephi, the same dataset yielded below shown graph containing 12 clusters. Fig 5 shows the communities formed and the percentage distribution of each community based on the number of nodes.

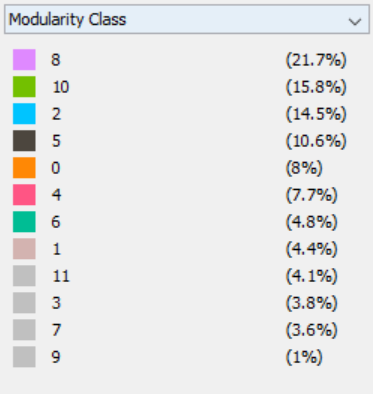


Fig 6 shows the different communities identified, each color representing a different community.

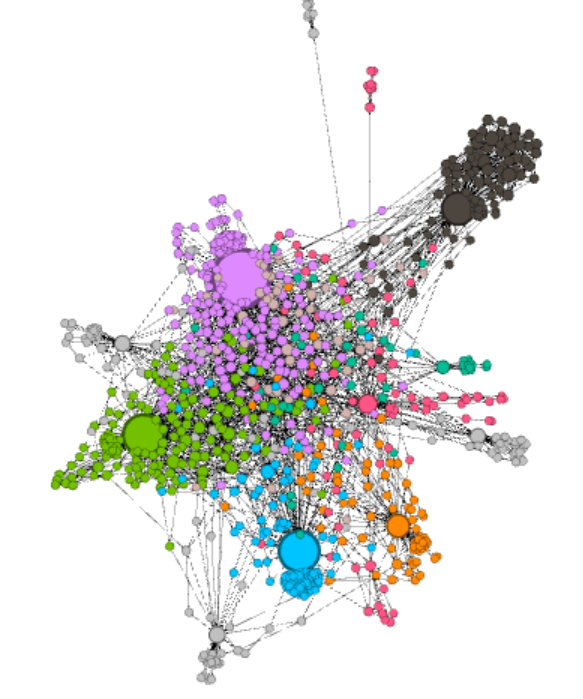


Fig 6: Community detection using Gephi

We also ran the same walk trap algorithm in R and identified the different communities. We identified 17 different communities in R. We then ran the Community significance test to differentiate the significant communities based on the P-value and identied 12 significant communities. Below are the screenshots of 3 such significant communities:

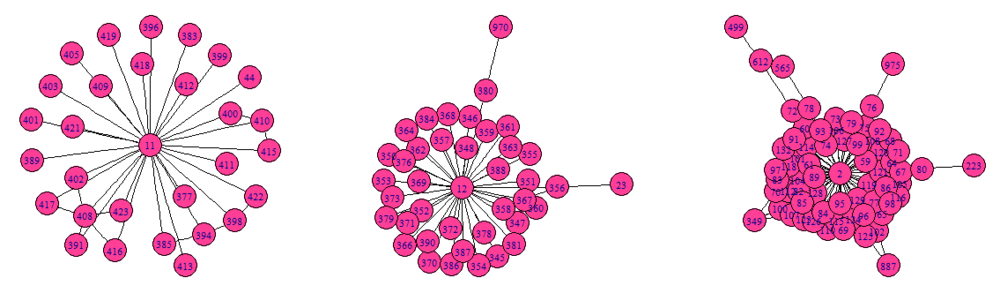


Fig 7: Different communities in R

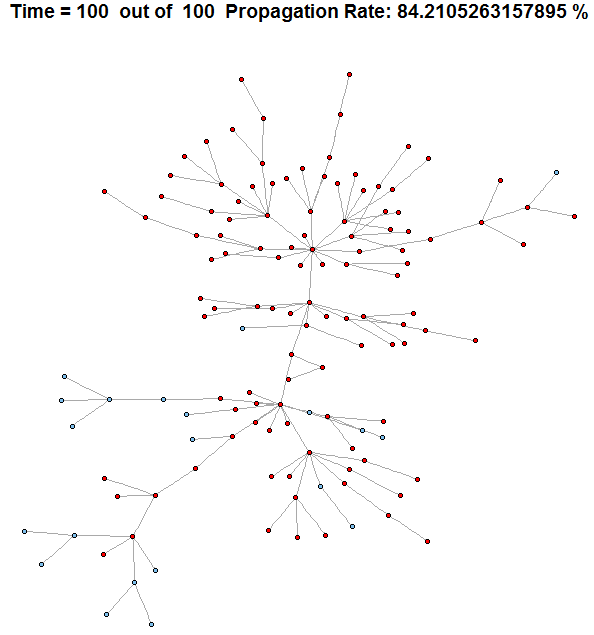
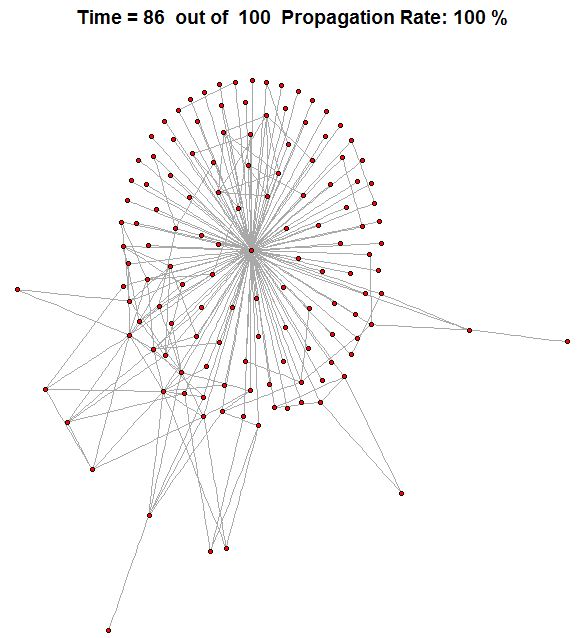
Epidemic Modeling

In order to analyze the flixster network and devise a suitable marketing/advertising campaign which addresses the following:

* *Target the right customers* – Ones who are most influential and targeting whom will lead to the awareness about the movie propagating to the whole community.
* *Advertising Campaign Duration* – Study the average number of time iterations or days it takes for the message to spread through the whole network.

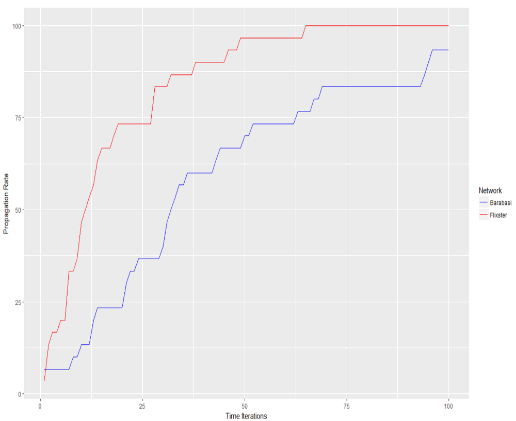
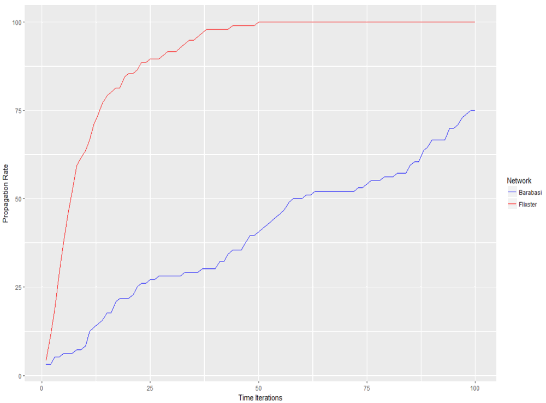
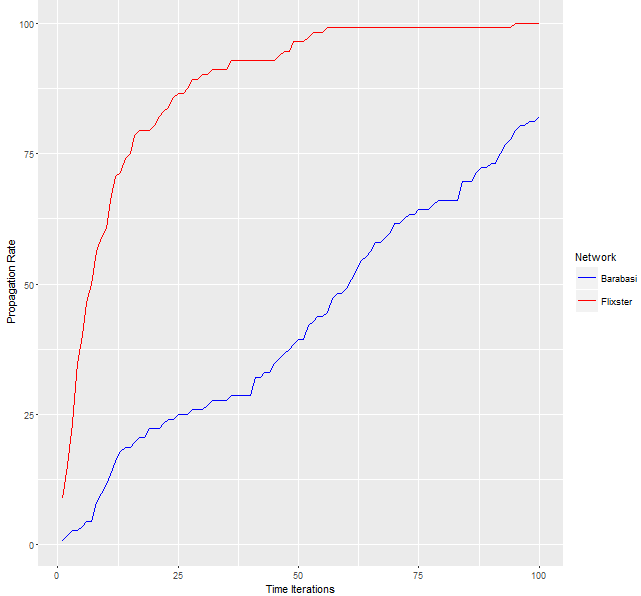
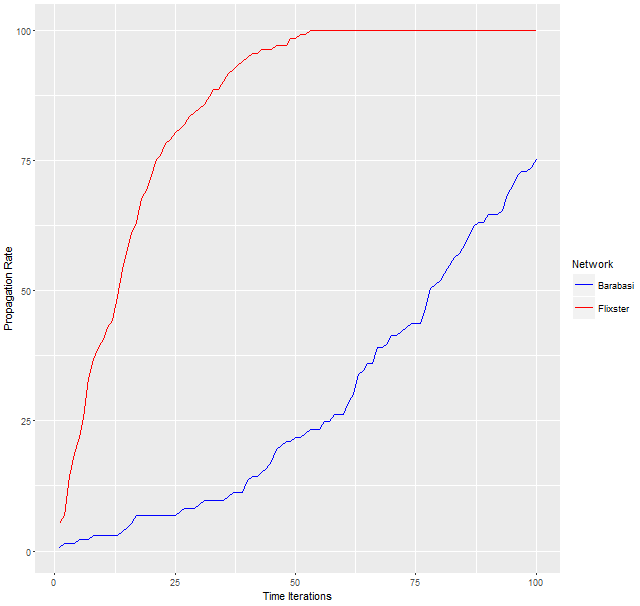
We intend to use the communities created earlier and study the spread of epidemic within each of the communities. Also, a Preferential Attachment Network of a similar size is obtained using the Barabasi.Game function and the infection rate in our flixster network is validated against this baseline. We have limited our analysis for 100 iterations for each of the communities and validated the actual infection rate at that point.

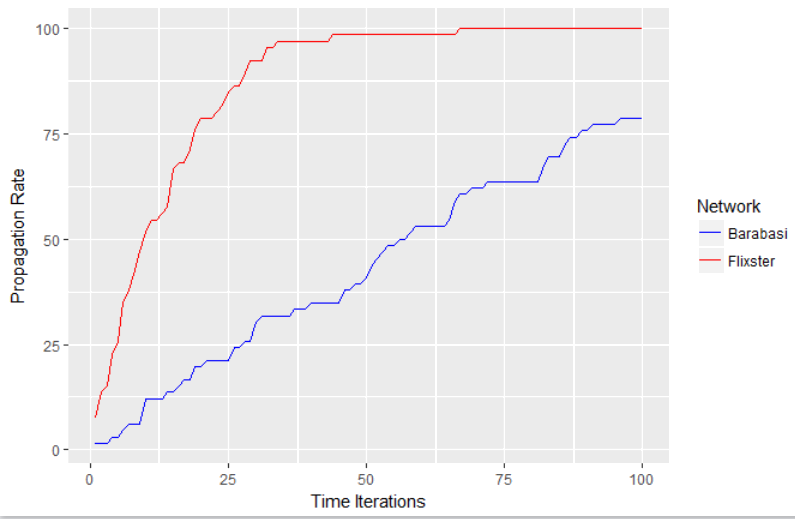
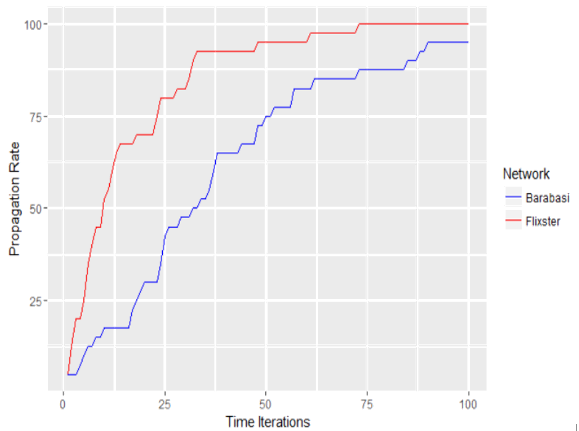
An example of the epidemic modelling on one of the communities is displayed below.



It can be noticed that the flixster network attains a higher infection rate much earlier than the Barabasi network of similar size. The epidemic modelling was repeated several times on the same community and a similar observation was made on each run.

A plot of the propagation rate in each of the selected communities against time for each of the communities is represented in the following plots.





A plot of the propagation rate shows that the increase in the flixster network is drastic initially and subsequently converges with minimal change. It can be noticed that after 50 time iterations or days, the increase in the awareness among the number of users is minimal. Whereas, in the barabasi network the growth in the infection within the community is much slower and in most cases do not attain a 100 percent infection before 100 iterations.