Social Network Analytics – Project Report

Social Network Analysis on Facebook Users

Team Members

Karthikreddy Kuna (kk3375@drexel.edu)
Pradeep Kankipati (pk593@drexel.edu)

Table of Contents:

Abstract: Social Network Analytics is one of the modeling techniques which illuminates the structure of the network and the importance of the individuals within the network. Network Centrality is a metric of the prominence of a node in a network, which allows revealing the structure patterns of the network. Community structure is an important property of network and with the increasing popularity, community structure is also getting equally complex within social media platforms like Facebook, Twitter, Instagram. Community detection algorithms try to find groups of nodes that have similarities. Link prediction is one of the most interesting network analysis methods which helps to predict potential associations to be formed in the future. This paper summarizes the different methods performed to find the influential person in Facebook users, how communities are formed based on the user and prediction of the future association of the users.

1. Introduction

The emerging growth and usage of social media platforms improve people's connectivity over the globe as a community. A community in a social network is referred to as a group of people who are tightly interconnected in the network. With this growth in the network of each individual and the information generated through it is unmanageable. So, community detection and influence propagations in social networks are widely studied because of its importance in uncovering how people connect, interact, and form social groups. However, little attention has been given to the community structure of social media users which bought our attention to analyze the way different communities are formed and how an individual acts as a key person in the transmission of information among the community and with other communities. In this project, we tried to analyze the Facebook social media data of various users and detect the closely-knit groups formed based on key user and predict the likelihood of friendship between two users in the future.

2. Literature Review

Previous research has stressed the need to identify the influential people within different organized groups. Social Network Analytics (SNA) can be used to identify the key nodes in the network using various centrality measures [1]. One of the most used centrality measures is degree centrality, that measures the number of direct neighbors connected to a node. The higher the degree centrality the more important the node to the network as node could represent a hub for information within the network.

Borgatti says that to measure centrality effectively, one should understand why centrality is important. He says we need to do this to maximize disruption to a given network and to maximize collection of information. In addition to this Borgatti recognizes that effective identification of centrality is done on multiple nodes in the network rather than central node [2]. As the research continues time to time, centrality measures like betweenness [3], closeness [4], and eigenvector centrality [5] were introduced to compute nodes importance in the network.

In the field of science, communities can be considered as subgraphs of a network. The whole network graph is a combination of several subgraphs and connections among the sub-graph nodes are intradense. Whereas nodes that belongs to different communities are sparse. Newman termed this sub-graph as community structure [6]. M. Girvan and M. E. J. Newman proposed community structure and detection algorithm in social and biological networks. Social groupings in a social network represents communities [7].

Later, Newman proposed modularity based fast greedy algorithm which improves the execution speed of GN. In a further study, Clauset et al pointed out that the update of the adjacency matrix in Newman's algorithm involves a great number of useless operations. They utilized a complex data structure, like the largest heap and balanced binary trees, reducing the complexity of the Fast-greedy algorithm [8].

In the research of social network analytics, Link prediction is one of the most interesting topics. Which exploits existing the network information like characteristics of the nodes and edges, to predict the potential links to be formed in the future. Link prediction can be done by several methods and we follow Local similarity measures in the prediction of the future association [9][10].

3. Research Questions/Problem Statement

This project deals with the ego network of Facebook user data and mainly focuses on analyzing the different communities of Facebook social network and social connectedness of different users. Below are the observations we analyzed on the Facebook network.

- Find the most influential person/ centrality of a social network.
- Identify tight communities of a given single user and his social network.
- Prediction of likelihood of future association of two nodes.

4. Ego Network

SNAP (Stanford Network Analysis Project), introduced the concept of viewing users as individual "Egos". They formulated the problem of circle detection as a clustering problem on the user ego-network, the network of friendships between user friends.

The Ego network is defined as a network of friendships between a user and his friends. In the below ego network, the user is referred to as the 'ego' and the nodes in this network are 'alters'. Circles are formed by densely connected sets of alters. Each circle is not only densely connected but its members also share common features.

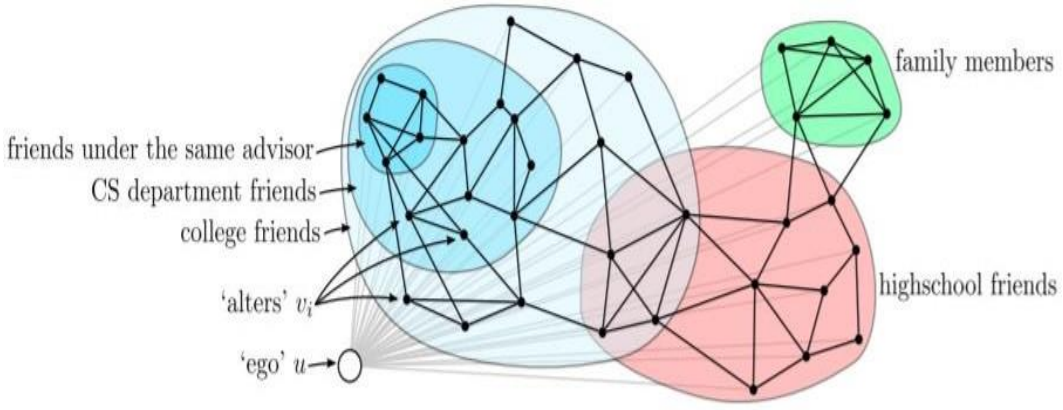

Fig. Ego network example

5. Dataset Description

We used the Facebook dataset acquired from https://snap.stanford.edu/data/ego-Facebook.html. This dataset consists of 'circles'/ 'Friend lists' from Facebook. Facebook data was collected from survey participants using the Facebook app. Facebook data has been anonymized by replacing the Facebook-internal IDs for each user with a new value. The data represents multiple users identified with a node_id. Each of these users is called an 'ego' and each ego consists of 5 files:

- *Node_id.circles* which represents the ego's set of circles as circle name and a series of node ids. Each
  record in file denotes a circle.
- *Node_id.edges* gives the connection between the alters of the ego node. The ego node is assumed to follow every node_Id present in the file.
- *Node_id.featname* provides the names of each feature of the ego node.
- *Node_id.egofeat* contains a series of 0's and 1's that represents if the ego user had specified his
  features on Facebook.
- *Node_id.feat* contains 0, 1 series of feature details for every alter of the ego user.
- *Facebook_combined* file which has node ID pairs denoting connections of all the ego users who are connected on Facebook instead of focusing on one ego. It is basically a list of all the edges on the network.

Below figure provides the statistics of the dataset.

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

Fig. Dataset Statistics

6. Dataset and Network Analysis

We used the Facebook combined file for our analysis and built the undirected graph for the whole network involving all the ego users using Networkx package. Networkx is a python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex network. We used functions from the networkx package to analyze the connection between the nodes, to find out the most influential person in the network using centrality measures and to analyze the formation of communities based on a user.

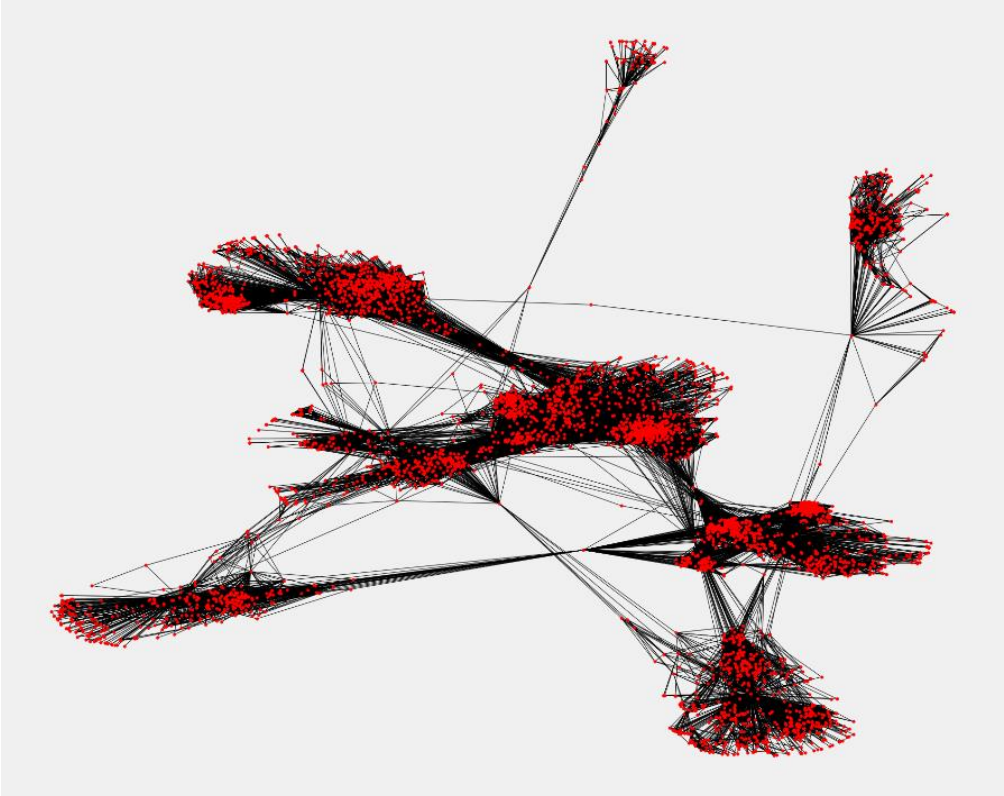

Fig. Network graph on Facebook Users

- **Nodes:** The individual users whose network we are building are referred as nodes.
- **Edges:** The connection between the nodes. It represents a relationship between the nodes within the network.

```
Number of nodes: 4039
Number of edges: 88234
```

- **Degree of the network:** Degree is the number of edges connected to a node. Average degree of all the node indicates the overall connectiveness of the network.

```
Average degree:  43.6910
```

- **Network Connectivity:** Checking the fully connectiveness of the Network built.

```
Connected Graph: True
```

- **Network Diameter:** The diameter of a network is the length of the longest geodesics in the network.

```
Diameter: 8
```

- **Clustering Coefficient:** Usually in a social network, users tend to form associations based on the connections they share. In other words, there is a tendency in a social network to form clusters. Clustering coefficient is the fraction of pairs of the node's connections that are connected to each other.

```
Average Clustering: 0.6055467186200876
```

6.1.1. Sub-Graph Generation:

Sub-graph was generated with criteria that users having minimum degree of 50 or above. We did this to validate the centrality metrics of original graph and the sub-graph. So, we can choose the most influential person in the network by comparing the metrics of both graphs.

```
Number of nodes: 1143
Number of edges: 50324
Average degree:  88.0560
```

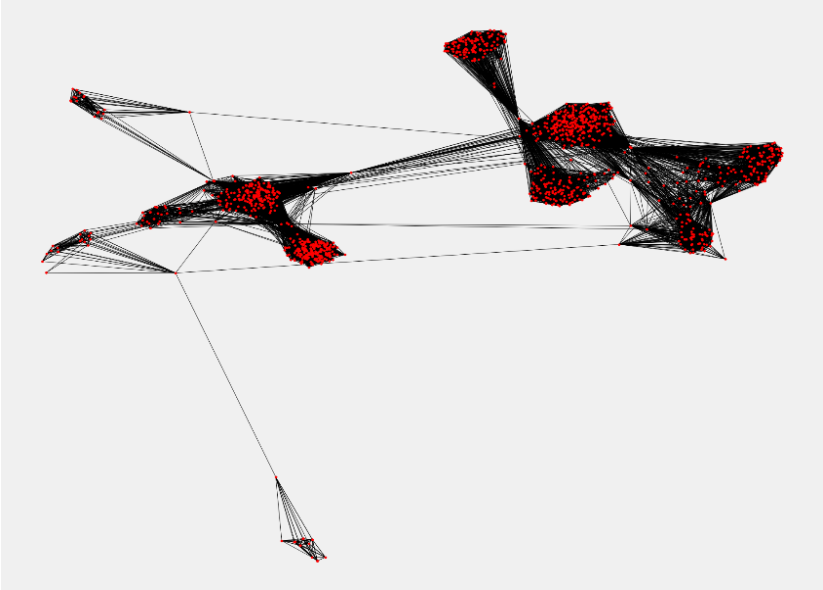

Fig. Sub-graph with minimum node degree 50

7. Analysis on Network Influencers

Centralities provide information pertaining to the most influential actors in the network. These parameters are also known as Centrality measures. Influential/ central nodes are the ones with a good range of connections with other nodes within the network. Centrality measures can help in identifying the most popular, liked, and biggest influencers within the network. In this project, we mainly analyzed the network using four major centrality measures for the Facebook network graph and sub-graph. Namely, Degree Centrality, Closeness Centrality, Betweenness Centrality, and Eigenvector Centrality.

7.1.1. Degree Centrality

The degree of an actor is defined as the number of connecting ties an actor has with others. To calculate the Degree centrality of an actor, the degree of an actor is divided by number of other actors in the network(n-1). Degree centrality metric defines the importance of an actor in a network as being measured based on its degree that is the higher the degree of an actor, the more important it is in a network.

Top 5-degree centrality nodes of Facebook user graph are [107, 1684, 1912, 3437, 0]

Top 5-degree centrality nodes of Facebook user sub graph are also [107, 1912, 2347, 2543, 1684]

Fig. Facebook User graph based on Degree centrality

Here we can observe degree centrality distribution for the Facebook user graph and sub-graph varied for top 5 nodes, which is 3437, and 0 nodes in the Facebook user graph are not present in the subgraph. In Facebook user graph 107 is the node with the highest degree centrality of 1045. Followed by nodes 1684,1912,3437,0.

7.1.2. Closeness Centrality

The closeness centrality metric defines the importance of an actor in a network as being measured by how close it is to all other actors in the network. For an actor, it is defined as the average of the geodesic distance between that actor to all other actors in the network. Closeness centrality is based on the truth that a node with the shortest distance will have more connectedness in the network.

Top 5-closeness centrality nodes of Facebook user graph are [107, 58, 428, 563, 1684]

```
[(107, 0.45969945355191255),
 (58, 0.3974018305284913),
 (428, 0.3948371956585509),
 (563, 0.3939127889961955),
 (1684, 0.39360561458231796)]
```

Top 5-closeness centrality nodes of Facebook user sub graph are also [107, 1577, 1718, 428, 1465]

```
[(107, 0.5282146160962072),
 (1577, 0.45716573258606885),
 (1718, 0.45425616547334924),
 (428, 0.4519192718638702),
 (1465, 0.4476675813406507)]
```

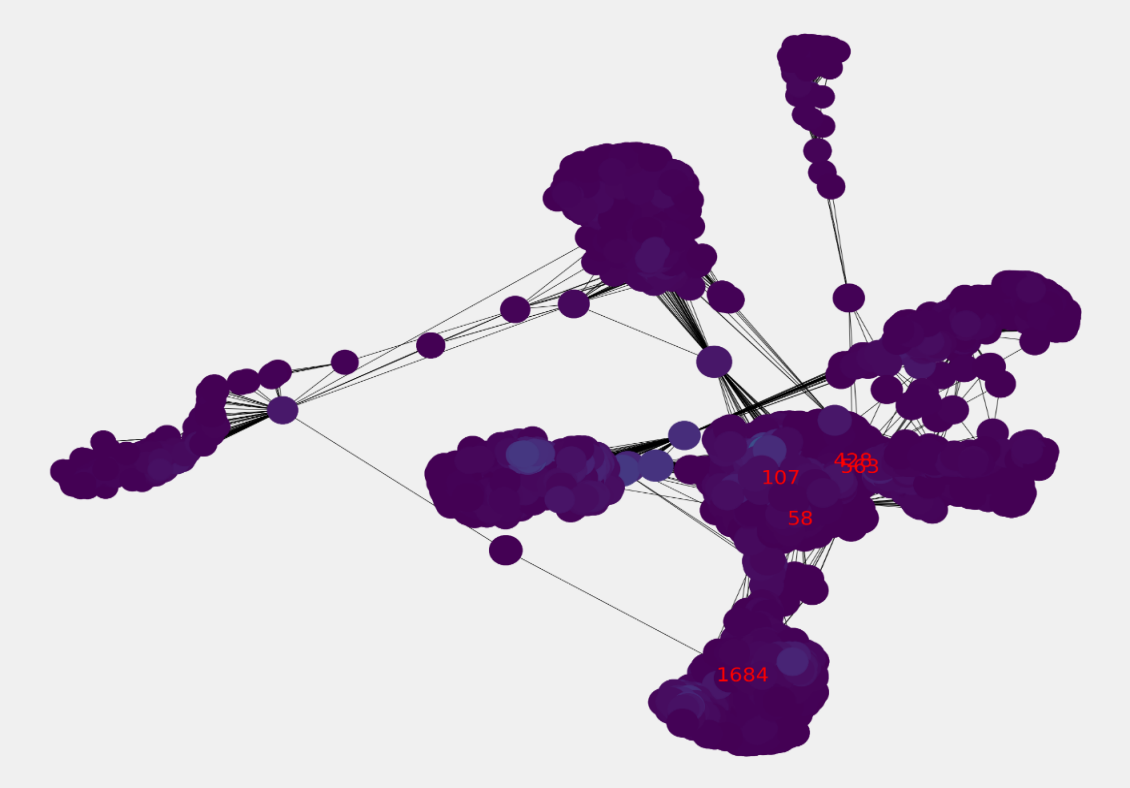

Fig. Facebook user graph based on Closeness centrality

We can observe that Closeness centrality varied from the Facebook user graph and Sub-graph for the top 5 nodes. Nodes 107 and 428 are present in the top 5 closeness centrality nodes in both graphs. Node 107 with the highest closeness centrality also has the highest degree centrality as well.

7.1.3. Betweenness Centrality

Betweenness centrality metric defines and measures the importance of an actor in a network-based upon how many times it occurs in the shortest path between all pairs of actors in the network. The actor with the highest betweenness centrality acts like a bridge between two social groups and plays a prominent role in the communication of information within the network. The nodes with high betweenness centrality can have more strategic control over other nodes, which help the node to influence the whole group.

Top 5-Betweenness centrality nodes of Facebook user graph are [107, 1684, 3437, 1912, 1085]

```
[(107, 0.48077531149557645),
 (1684, 0.33812535393929544),
 (3437, 0.23649361170042005),
 (1912, 0.22967697101070242),
 (1085, 0.14943647607698152)]
```

Top 5-Betweenness centrality nodes of Facebook user sub graph are also [107, 1684, 1912, 1718, 1577]

```
[(107, 0.5530400036447516),
 (1684, 0.346713961884091),
 (1912, 0.1867928883663688),
 (1718, 0.13945898618781966),
 (1577, 0.12701324126479357)]
```

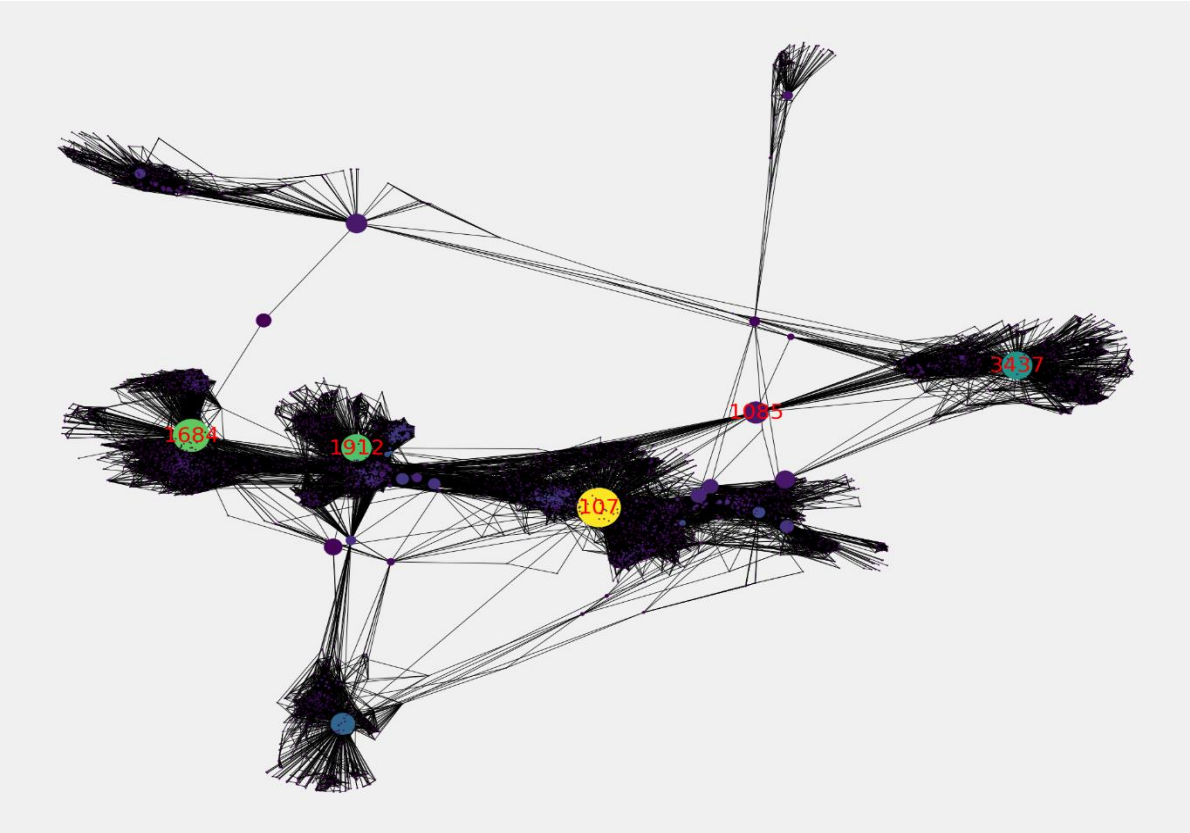

Fig. Facebook user graph based on Betweenness centrality

Here we can observe that Node 107 has the highest betweenness centrality in both the Facebook user graph and sub-graph. It is interesting to note that node 107 is the top in degree, betweenness, closeness centrality measures, and node 1684 also in the top 5 nodes of three metric calculated. Node 1912 is in the top 5 nodes when performing degree, betweenness, and closeness centrality measures.

7.1.4. Eigenvector Centrality

Eigenvector centrality is a measure of the influence of an actor in a network. It assigns relative scores to all actors in the network based on the concept that connections to high-scoring actors contribute more to the score of the actor in question than equal connections to low-scoring actors. In other words, Eigenvector centrality decides the importance of a node based on the importance of the other connected nodes. Google's Pagerank algorithm is a variant of the Eigenvector centrality algorithm.

Top 5-Eigenvector centrality nodes of Facebook user graph are [1912, 2266, 2206, 2233, 2464]

```
[(1912, 0.09540696149067629),
 (2266, 0.08698327767886553),
 (2206, 0.08605239270584343),
 (2233, 0.08517340912756598),
 (2464, 0.08427877475676092)]
```

Top 5-Eigenvector centrality nodes of Facebook user sub-graph are [1912, 2266, 2206, 2233, 2464]

```
[(1912, 0.09209797568781998),
 (2266, 0.08634034441692785),
 (2206, 0.08624255422568931),
 (2233, 0.08518486431929224),
 (2464, 0.08461774213989234)]
```

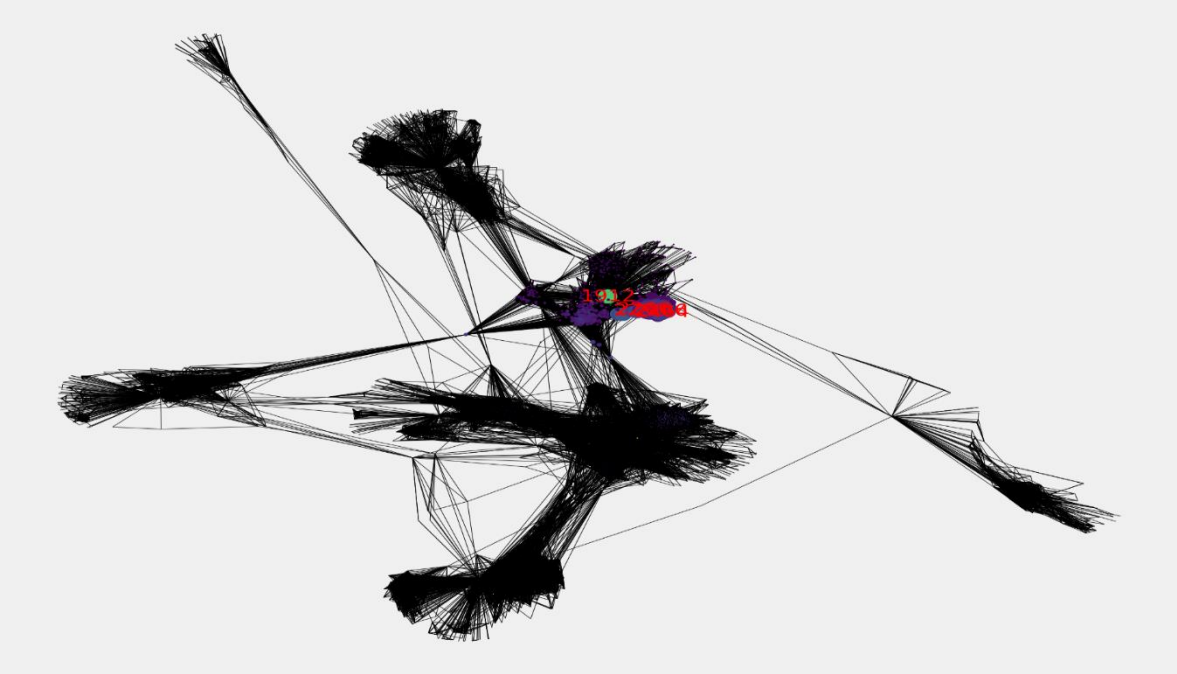

Fig. Facebook user graph based on Eigenvector centrality

We can observe that the top 5 nodes in both the Facebook user graph and the sub-graph are the same. Node 1912 is the top node in Eigenvector centrality measure followed by 2266, 2206, 2233, 2464. Node 1912 is in the top 5 nodes in eigenvector, degree and betweenness centralities.

The table below is the lists of the top 5 nodes in Centrality measures on Facebook user graph.

Degree Centrality	Closeness Centrality	Betweenness Centrality	Eigenvector Centrality
107	107	107	1912
1684	58	1684	2266
1912	428	3437	2206
3437	563	1912	2233
0	1684	1085	2464

The table below is the lists of the top 5 nodes in Centrality measures on Facebook user sub-graph.

Degree Centrality	Closeness Centrality	Betweenness Centrality	Eigenvector Centrality
107	107	107	1912
1912	1577	1684	2266
2347	1718	1912	2206
2543	428	1718	2233
1684	1465	1577	2464

We can observe that few nodes are common between Degree Centrality, which is the measure of degree, and Betweenness Centrality which controls the information flow and Eigenvector centrality which is calculated based on the importance of the nodes within a network. It is obvious that important nodes that are connected also lie on the shortest paths between other nodes. So, according to the centrality measures we considered, we choose node **'1912'** as the most influential/ centrality node in the Facebook user network, followed by **'107'**.

8. Analysis on Community Detection

We want to perform the analysis of the tight communities formed based on key nodes. To achieve this, we have created sub-graphs of nodes '1912' and '107'. Subgraph was created based on the concept of neighborhood and connectedness. We further analyze the identification of the communities and then discuss their robustness using the Modularity. We have implemented the Girvan-Newman algorithm which was based on edge betweenness of nodes and analyzed how tightly knit communities were formed. We also performed community detection analysis using Clauset-Newman-Moore- fast greedy algorithm from the networkx package.

8.1.1. Subgraph – 1912 Generation

Subgraph of node 1912 was generated based on neighbors around the key node 1912.

```
Number of nodes: 1003
Number of edges: 33999
Average degree:  67.7946
```

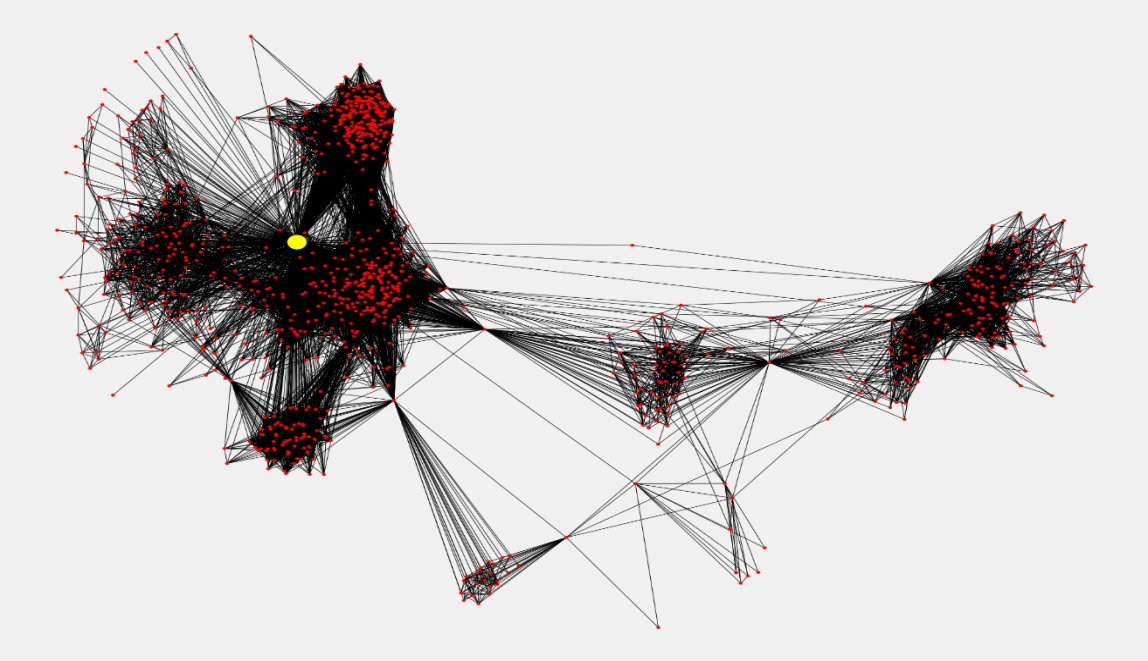

Fig. Subgraph 1912

8.1.2. Subgraph - 107 Generation

Subgraph of node 107 was generated based on neighbors around the key node 107.

```
Number of nodes: 2687
Number of edges: 58061
Average degree:  43.2162
```

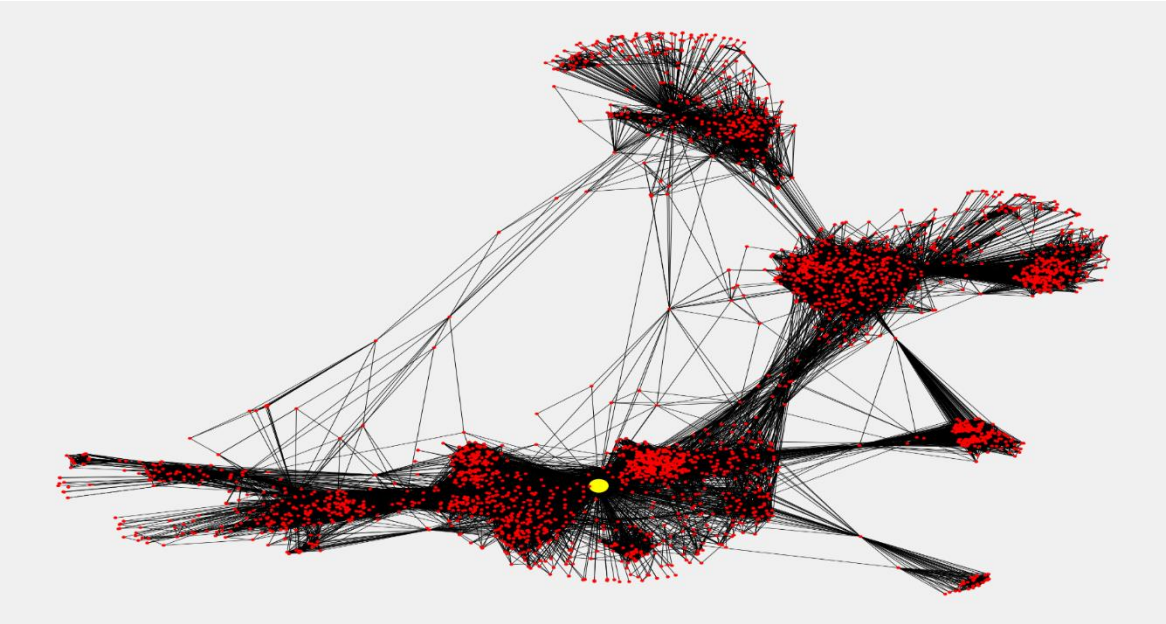

Fig. Subgraph 107

8.2. The Girvan-Newman algorithm

The Girvan-Newman algorithm detects communities by removing edges with the largest edge betweenness one by one from the network graph. The network will be divided into connected components. The connected components of the remaining network are the communities. The Girvan-Newman algorithm focuses on edges that are most likely between communities. If a graph has communities that are only loosely connected by groups with less interconnected edges, then all shortest paths between communities must be in the path of these edges. So, we can say that edges with high edge betweenness will connect the communities. The major impediment in following the Girvan-Newman algorithm is with the time parameter. Processing time for this algorithm lasts for hours in a few graphs. We detected communities on Facebook user graph, sub-graph, user 1912 sub-graph, user 107 sub-graph using this algorithm, and evaluated the modularity of each graph communities.

8.2.1. Girvan-Newman on Facebook user graph

Two communities were detected with modality value 0.043. We can further divide the communities into sub-communities.

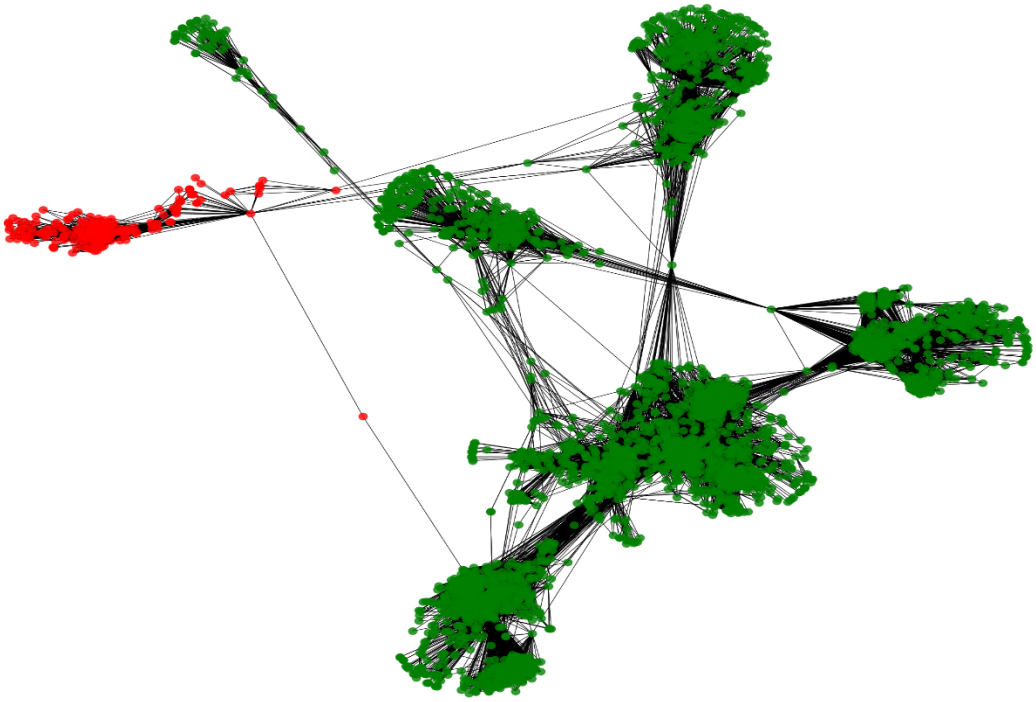

Fig. Communities detected on Facebook user graph

8.2.2. Girvan-Newman on Facebook user sub-graph

A total of 5 communities was detected with modality value 0.601

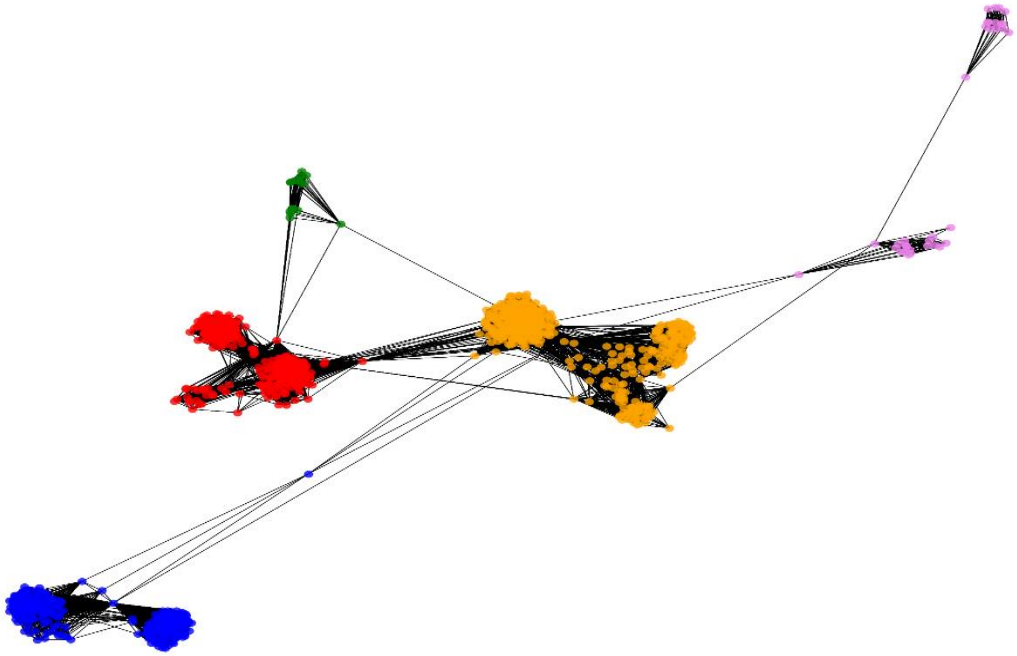

Fig. Communities detected on sub-graph

8.2.3. Girvan-Newman on Facebook user 1912 sub-graph

A total of 5 communities was detected with modality value 0.167

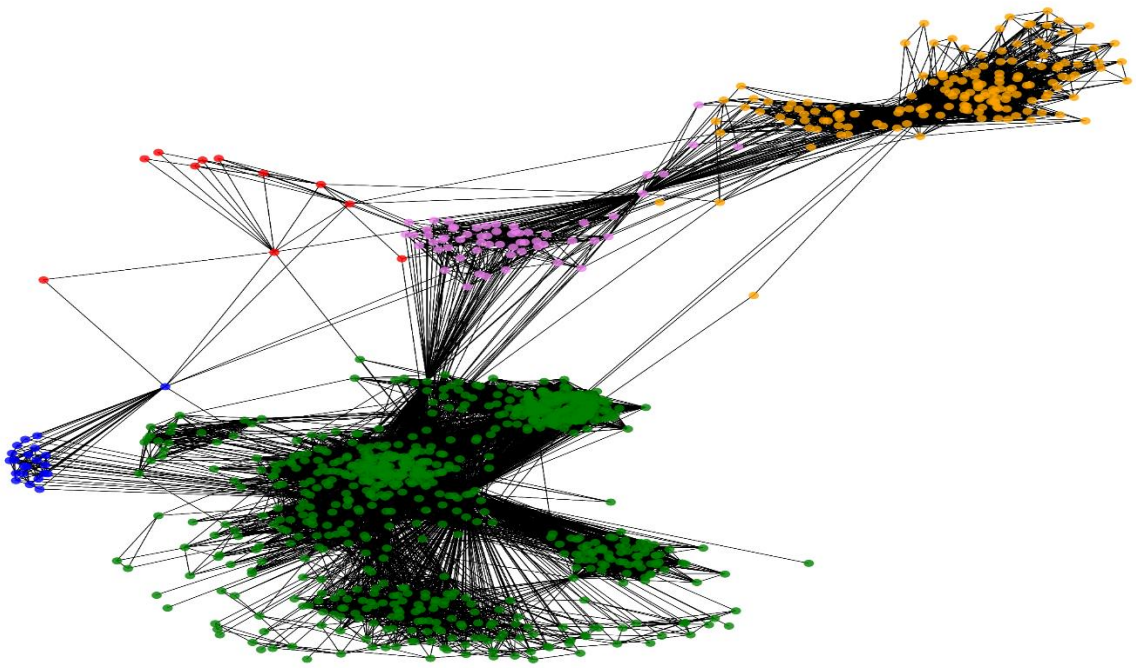

Fig. Communities detected on User 1912 Subgraph

8.2.4. Girvan-Newman on Facebook user 107 sub-graph

Two communities were detected with modality value 0.092. We can further divide the communities into sub-communities.

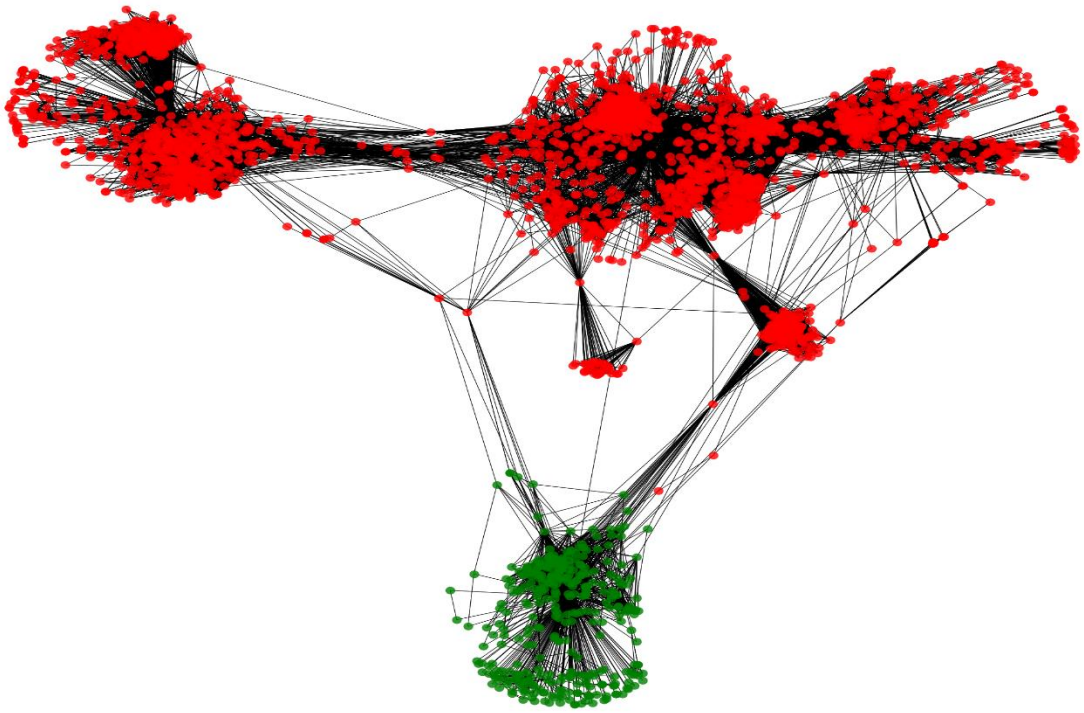

Fig. Communities detected on User 107 Subgraph

8.3. Greedy Modularity Algorithm

In the Greedy Modularity algorithm, Communities identification within the network can be done by optimizing modularity measurement which is denoted by Q. It indicates how well the clusters are obtained from the overall network. Ideally, nodes within the community should have more connections within the community rather than having connections outside of the community. The community is said to be stronger if it has a larger Q value. Greedy Modularity algorithm is much faster than the Girvan-Newman algorithm, communities were detected in few seconds while Girvan-Newman took a few hours. We detected communities on Facebook user graph, sub-graph, user 1912 sub-graph, user 107 sub-graph using the Greedy Modularity algorithm, and evaluated the modularity of each graph communities.

8.3.1. Greedy Modularity on Facebook user graph

A total of 13 communities was detected with modality value 0.777

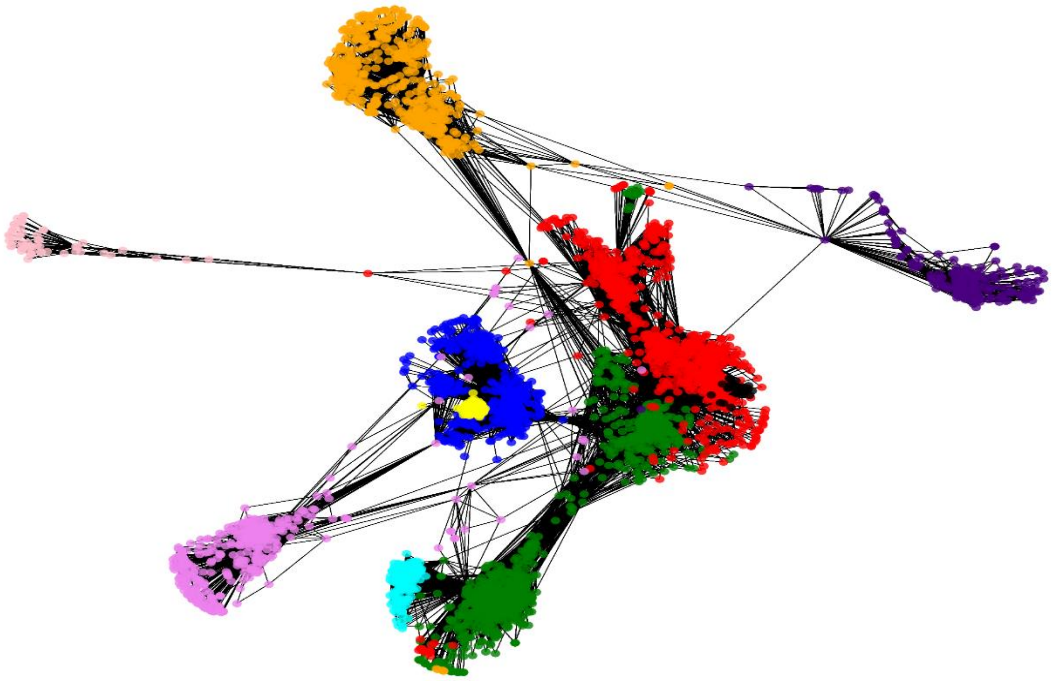

Fig. Communities detected on Facebook user graph

8.3.2. Greedy Modularity on Facebook user sub-graph

A total of 8 communities was detected with modality value 0.736

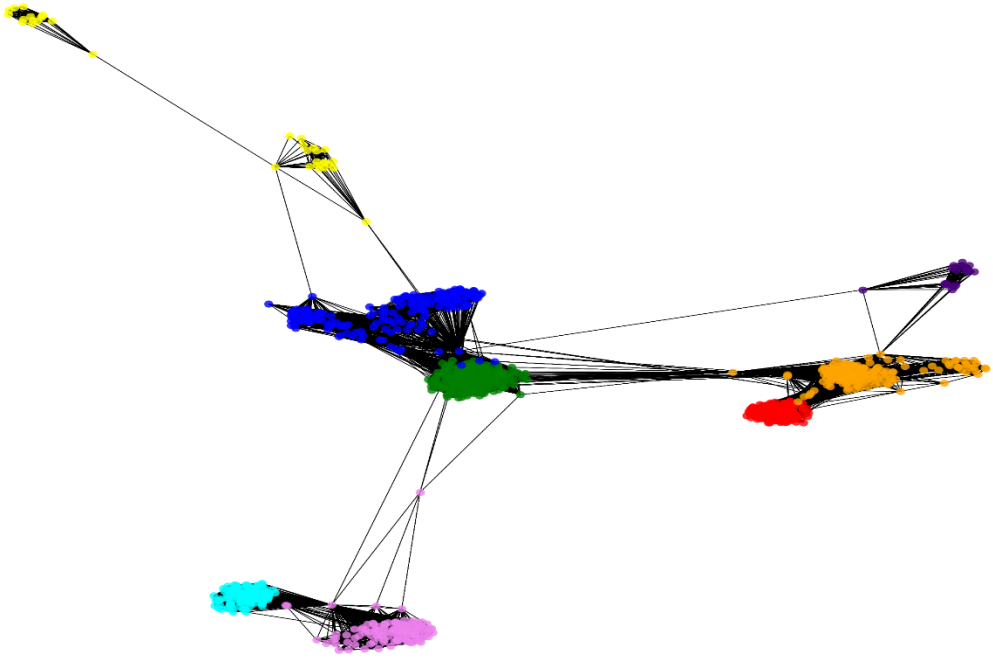

Fig. Communities detected on Subgraph

8.3.3. Greedy Modularity on Facebook user 1912 subgraph

A total of 5 communities was detected with modality value 0.585

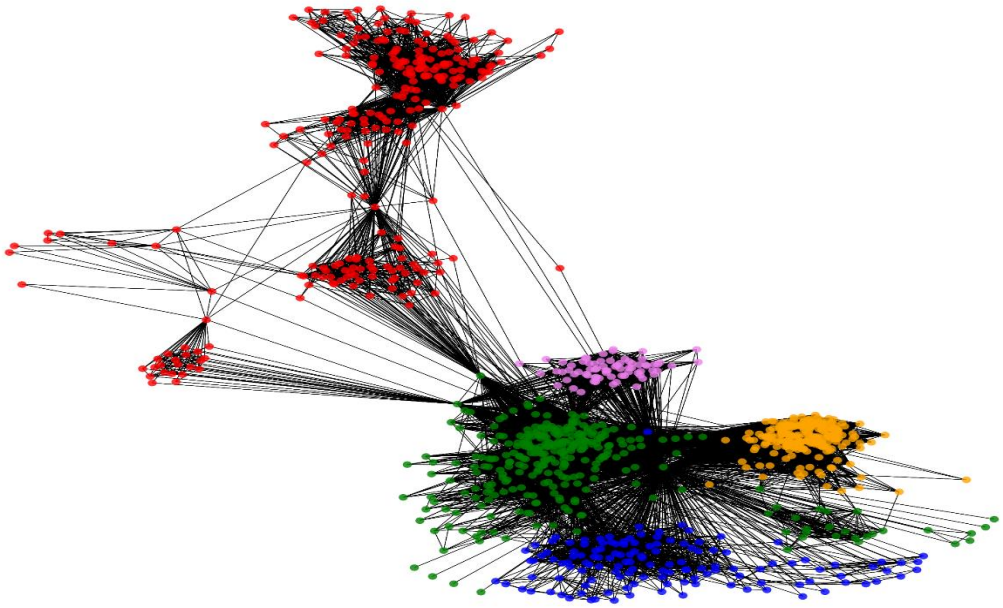

Fig. Communities detected on 1912 Subgraph

8.3.4. Greedy Modularity on Facebook user 107 subgraph

A total of 11 communities was detected with modality value 0.767

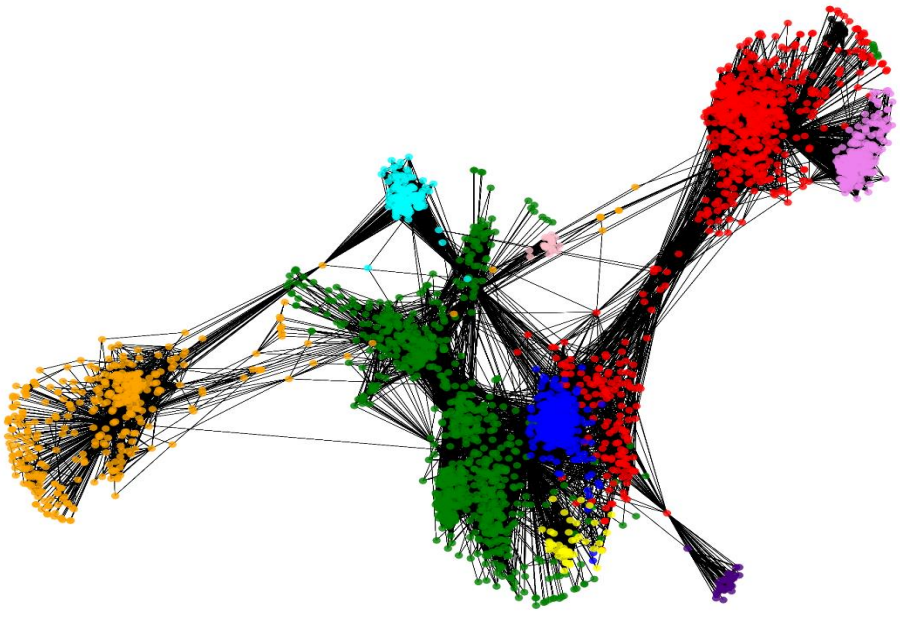

Fig. Communities detected on 107 Subgraph

8.4. Evaluation of community detection

We choose Modularity to measure community strength. Modularity exhibits a systematic tendency to h ave more intracommunity links and quantifies the extent to which the network forms the communities. Communities with higher modularity are known to have dense connections between the nodes within th e group and have sparse connections between nodes in a different group.

Graph	Modularity	
	Girvan-Newman	Greedy Modularity
Facebook User graph	0.043	0.777
Facebook User Sub-graph	0.601	0.736
Facebook User 1912 Sub-graph	0.167	0.585
Facebook User 107 Sub-graph	0.092	0.767

Based on the modularity, number of communities and time taken, fast greedy algorithm performed well when compared with Girvan-Newman.

9. Prediction of Likelihood of future association

Prediction of the future association between two nodes is one of the most prominent research topics in Social network analytics. The objective of this analysis is to predict the pairs of nodes that either associate with others in the future or not.

Firstly, we extracted positive and negative samples, Positive indicates that there is an association between the nodes, and negative indicates there is no association. To generate the missing edges, we randomly selected two nodes from the data and validated the shortest path between two nodes must be greater than two. Nodes with geodesics less than two, they are most likely to be friends with each other. So, we eliminated this case so that our model could perform better.

Next, the major task we have is feature extraction to a file. We used the feature engineering techniques like common neighbors, resource allocation index, Jaccard coefficient, Adamic-Adar index, Preferential attachment using networkx package.

- Common Neighbors is the number of common neighbors of two nodes.
- Jaccard Coefficient is the number of common neighbors normalized by the total number of neighbors.
- Resource Allocation index is the fraction of a resource that a node can send to another through their common neighbors.
- Adamic Adar index is defined as sum of the inverse logarithmic degree centrality of the neighbors shared by the two nodes.
- Preferential Attachments is a model works on the concept that nodes with highest degree get more neighbors.

Next, we generated the train and test data using the feature extracted data file. We did some research and found the Support Vector Machine (SVM) Model best suits to our analysis. So, we built SVM and Logistic regression models. We successfully built the models and the SVM model got an impressive accuracy of 0.957 whereas the Logistic regression model got an accuracy of 0.955.

```
SVM accuracy: 0.9572198581560284
SVM F1 Score: 0.9572196035178815

Logistic regression accuracy: 0.9553475177304964
Logistic regression F1 Score: 0.9553421071553178
```

10.Conclusion

To summarize the analysis, we performed a network analysis on the Facebook user graph and various sub-graphs that were created based on different centrality measures. We performed various centrality measures on the graph network and identified the influential nodes. We also built the Girvan-Newman algorithm and performed community detection using Girvan-Newman and fast greedy algorithms. We analyzed the communities formed and modularity of the network graphs. We then performed the link analysis on the future association of the graph, extracted the features using different feature engineering techniques, and build and evaluated the model using SVM and Logistic regression.

References:

[1] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications, M. Granovetter, Ed. Cambridge University Press, 1994.

[2] S. Borgatti, “Identifying sets of key players in a social network,” Computational and Mathematical Organization Theory, vol. 12, no. 1, pp. 21–34, 2006.

[3] L. C. Freeman, “A set of measures of centrality based on betweenness,” *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[4] L. C. Freeman, “Centrality in social networks conceptual clarification,” *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[5] P. Bonacich, “Factoring and weighting approaches to status scores and clique identification,” *The Journal of Mathematical Sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[6] M.E.J. Newman, "Detecting Community Structure in Networks", Eur. Phys. J. B 38, pp. 321-330, 2004.

[7] M. Girvan and M. Newman, "Community Structure in Social and Biological Networks", Proceedings of the National Academy of Scinces, vol. 99, no. 12, pp. 7821-7826, June, 2002.

[8] Clauset, A., Newman, M. E., & Moore, C. “Finding community structure in very large networks.” Physical Review E 70(6), 2004.

[9] A Clauset, C Moore and M E J. Newman, "Hierarchical structure and the prediction of missing links in networks[J]", *Nature*, vol. 453, no. 7191, pp. 98-101, 2008.

[10] Sucheta Soundarajan and John Hopcroft, “Using Community Information to Improve the Precision of Link Prediction Methods”

[11] J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.

[12] http://snap.stanford.edu/snappy/doc/reference/CommunityGirvanNewman.html