Network Science for the Web

Mini project assignment results

Group:

IMT2016057 Lakshmi Manonmaie
IMT2016072 Nomula Rohith Yogi
IMT2016110 Dachepalli Soumith Kumar
IMT2016078 Kalva Vaishnavi
IMT2016108 Arunaav

Train Bombing

This undirected network contains contacts between suspected terrorists involved in the train bombing of Madrid on March 11, 2004 as reconstructed from newspapers.

A node represents a terrorist and an edge between two terrorists shows that there was a contact between the two terrorists . The edge weights denote how 'strong' a connection was.

This includes friendship and co-participating in training camps or previous attacks.

64 vertices (terrorists)

243 edges (contacts)

Undirected and Weighted graph.

Degree Centrality

The terrorist with most number of connections.

By plotting we found out that the node with max centrality is corresponding to Jamal Zougam who proved to be main suspect in the attack.

Eccentricity

Inverse of max distance from that node to any other node i.e., it is near to all other nodes.

So the node with less eccentricity is the one who is connected to farthest node also better, as we can see Jamal is one in first few of who have the strong connection even with the last person
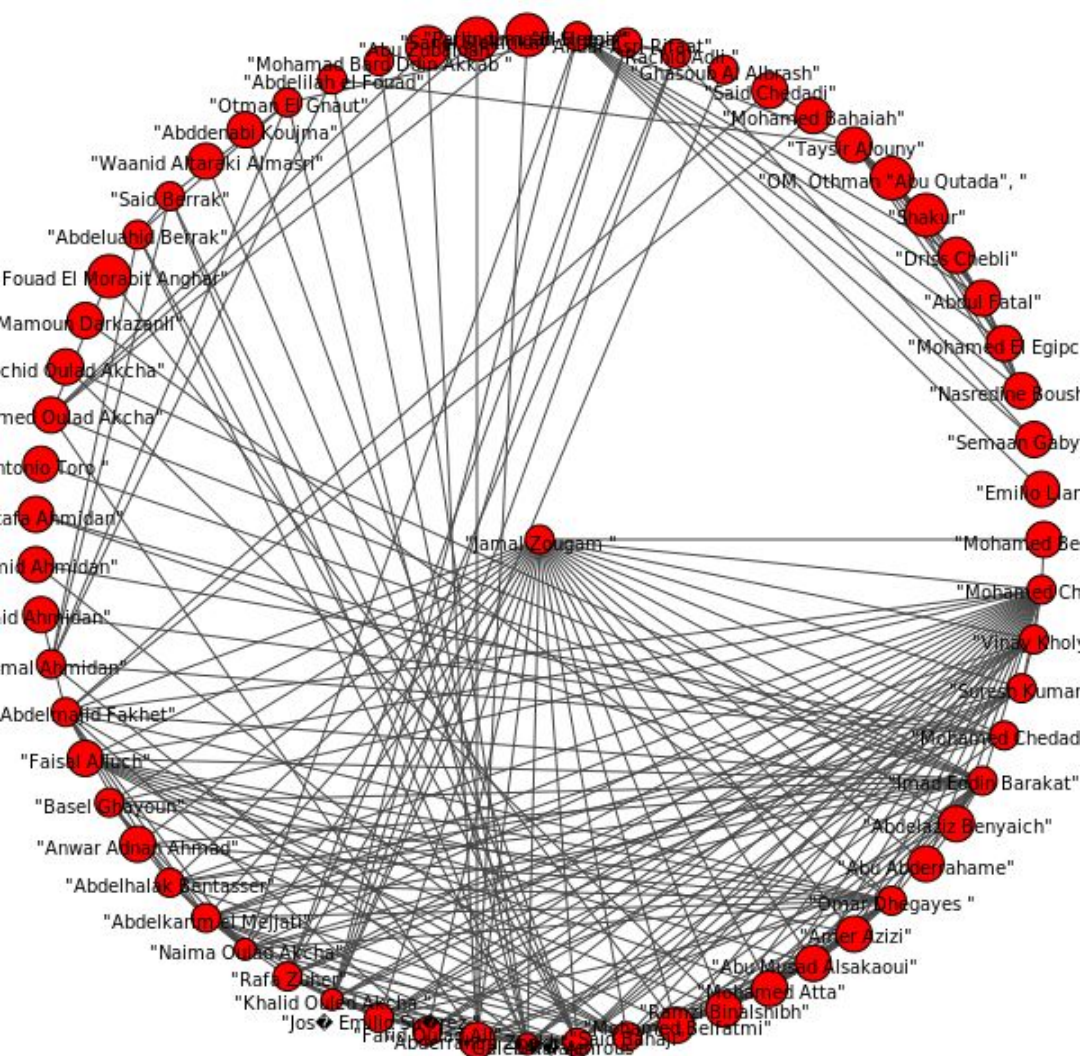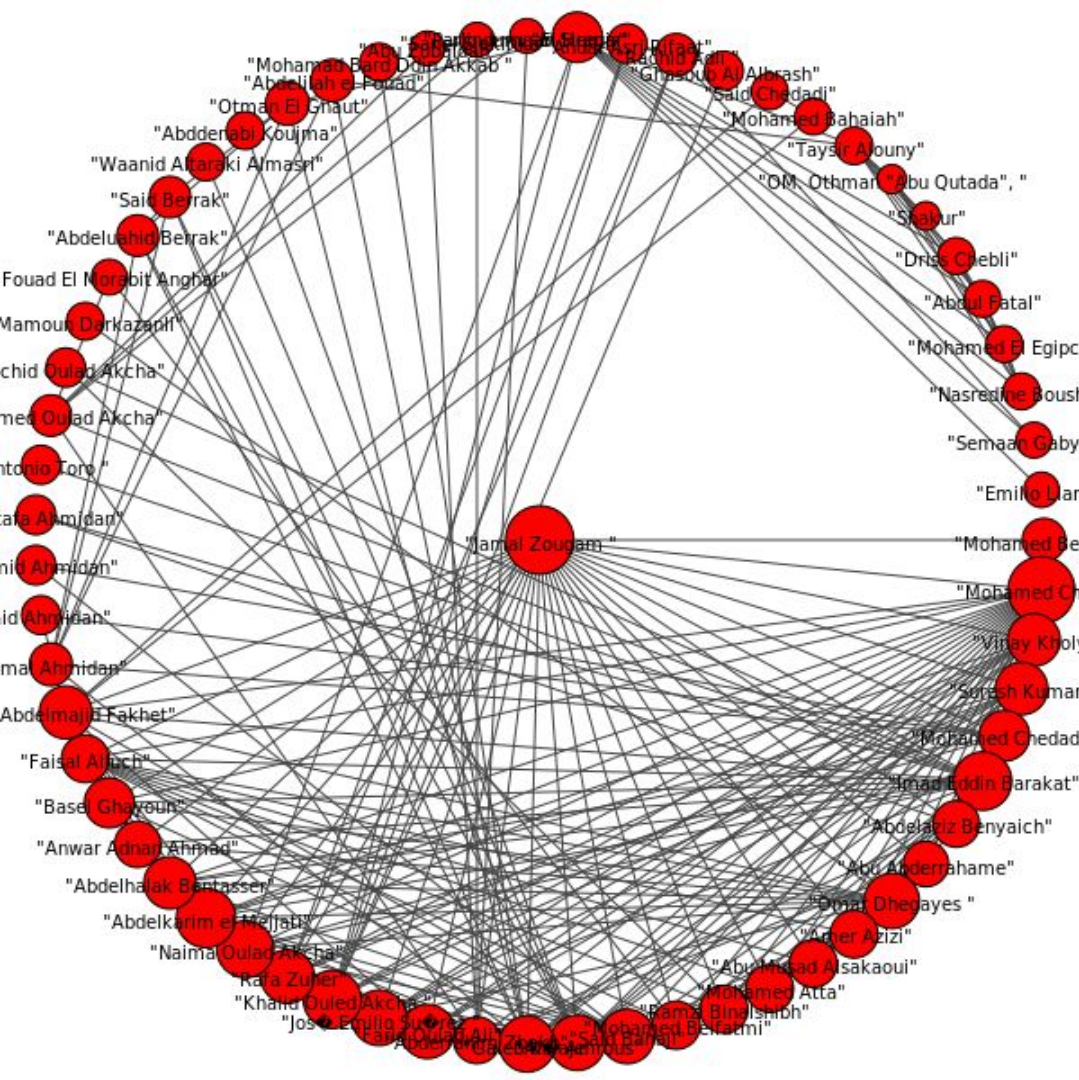
Closeness

Inverse of sum of distances to all the node.

The node which is close to most of the other nodes.

Compare with eccentricity.

Once again Jamal took his position in being the node with most closeness.

Radiality

Measure of how closely the node is integrated in the network (summation over one node(diameter-distance between 2 nodes)).

Almost every node is having same radiality, so we can conclude that almost every-one is closely integrated.

But here also Jamal is one with maximum radiality also.

Stress Centrality

Number of shortest paths that pass through that node.

If that node is removed many nodes will have to go in long path.

Here the maximum stress centrality is for Mohamed Chaoui who is the half brother to Jamal the main suspect of the attack.

So here we may think that lot of close relations to Jamal might be through Mohamed Chaoui.

Betweenness

Ratio of number of shortest paths between any 2 pairs of nodes passing through a node.

This centrality is high for El Hamir (some nodes are connected only through him)

Flow Vitality

This tells if a particular node is removed how much there will be decrease in flow. If a node is having high flow vitality it means removing that node will decrease the flow in the network. Maximum flow vitality is El Hemir.

Because by removing him the graph would be disconnected some of the terrorists are involved only through him. The flow vitality of Jamal is less because the connections he had are popular members connections, who are connected to lot other people also. So even if he is removed the information can be circulated among all.

Closeness vitality

Extent at which cost of all pairs of communication increases with the removal of a node.

Diff between Wiener index with and without that node.(Wiener index is summation of all pairs shortest paths).

This is maximum for Parlindumgan Siregar.

Stress Vitality

Shortest paths that have remained same before and after the removal of a particular node.

This measure of centrality will be more if the stress on that node is more.

So this is maximum for Mohamed Chaoui and can draw similar conclusions as stress of that node is maximum.

DataSet : Dolphins

This is a Undirected social network of bottlenose dolphins.

The nodes are the bottlenose dolphins of a bottlenose dolphin community.

An edge indicates a frequent association.

62 vertices (network is connected).

159 edges (associations)
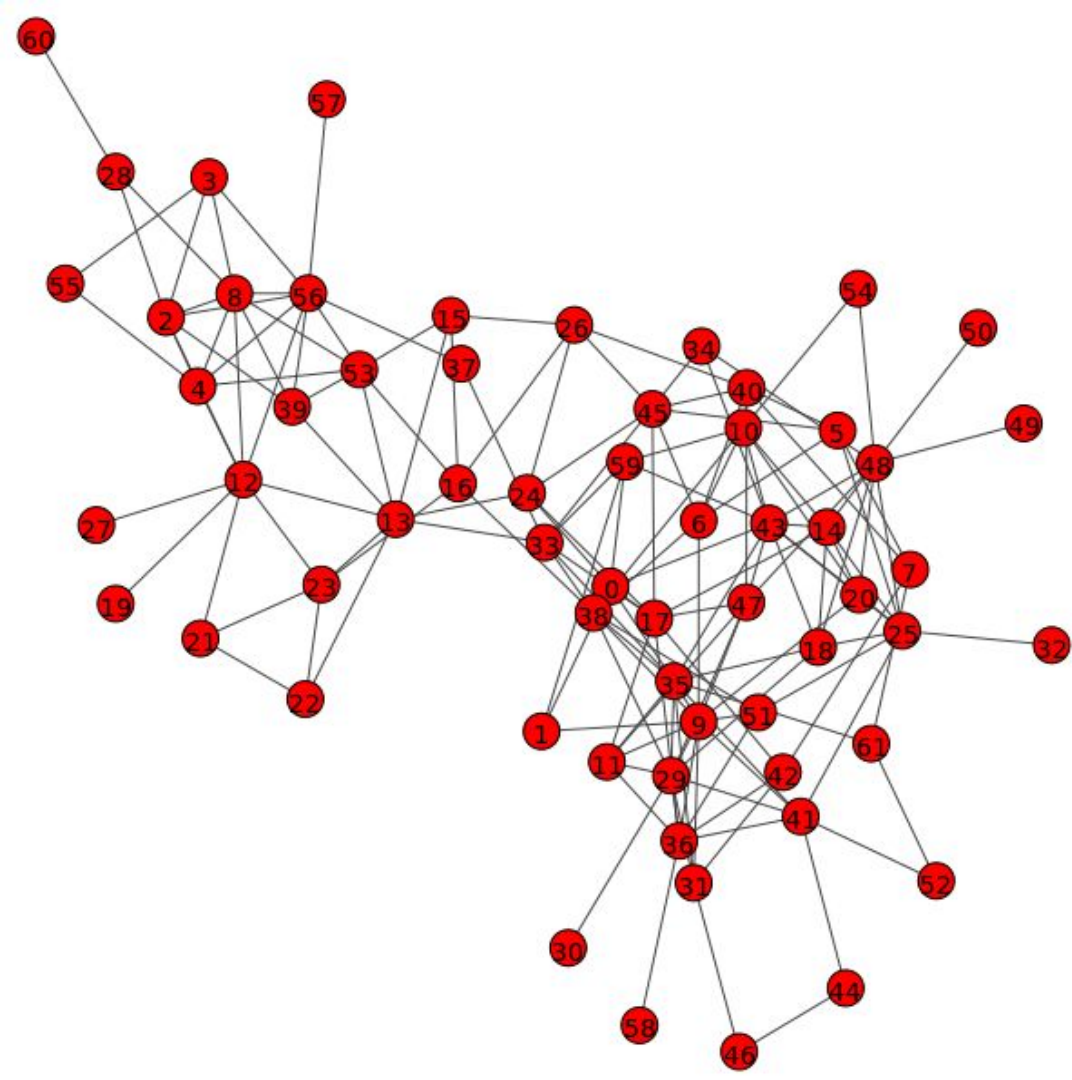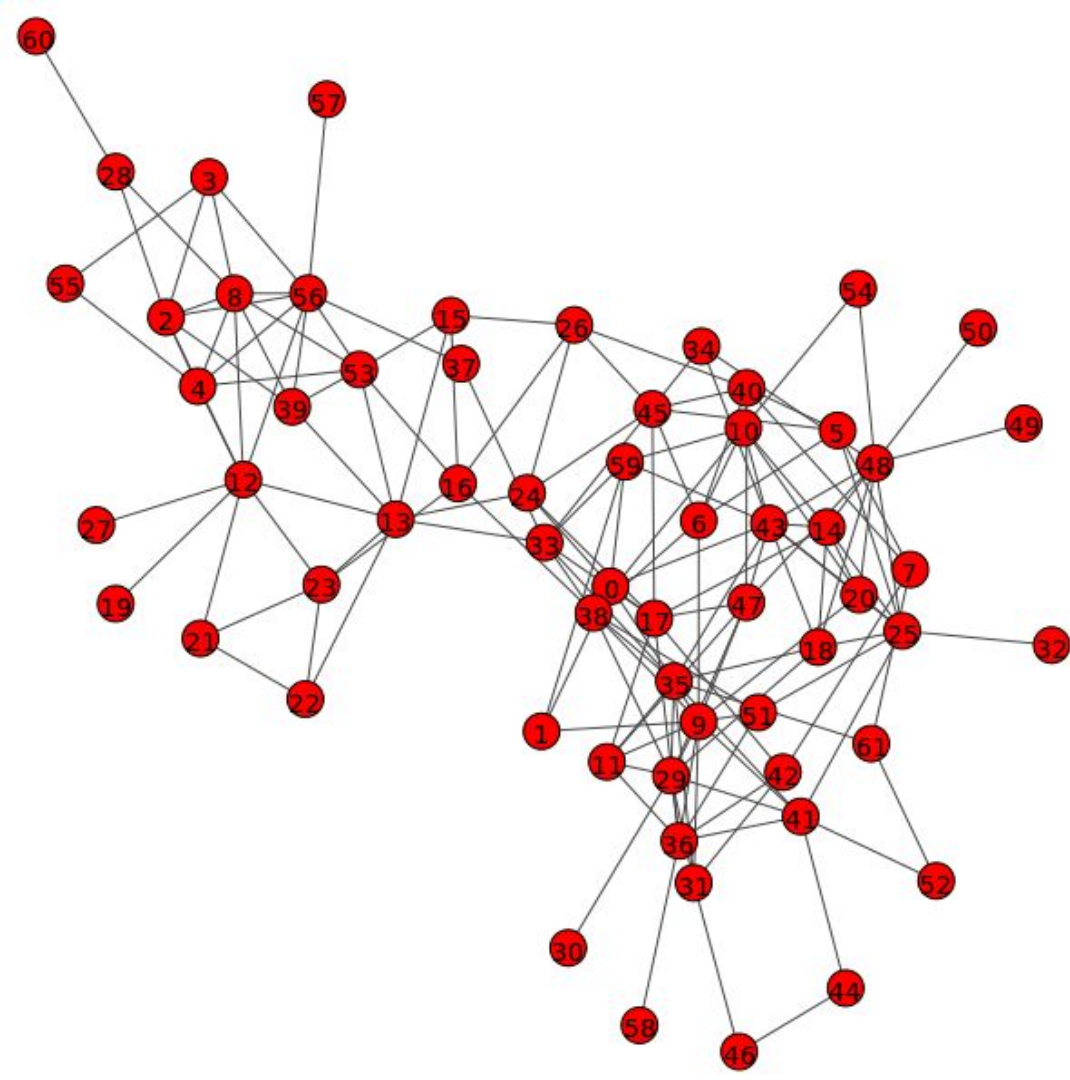
DataSet : Dolphins

Graph Property : **Weiner Index**

Connected So the Weiner Index is not infinity.

Given a graph G = (V,E) the Wiener Index is defined as the sum of all distances over all vertex pairs.

We used Networkx module to find out the Wiener Index of this graph.

nx.wiener_index(Gd)

Value : 6348.0

DataSet : Dolphins

Graph Property : **Max clique/s**

Also called as Perfectly dense groups.

Given a graph G = (V,E) a subset of nodes(U) is a Clique if G(U) is a complete graph.

Gd.largest_cliques()

returned : [(56, 8, 2, 4, 12),
         (48, 43, 14, 25, 20),
         (48, 43, 14, 25, 18)]

3 max cliques of size 5 were detected.
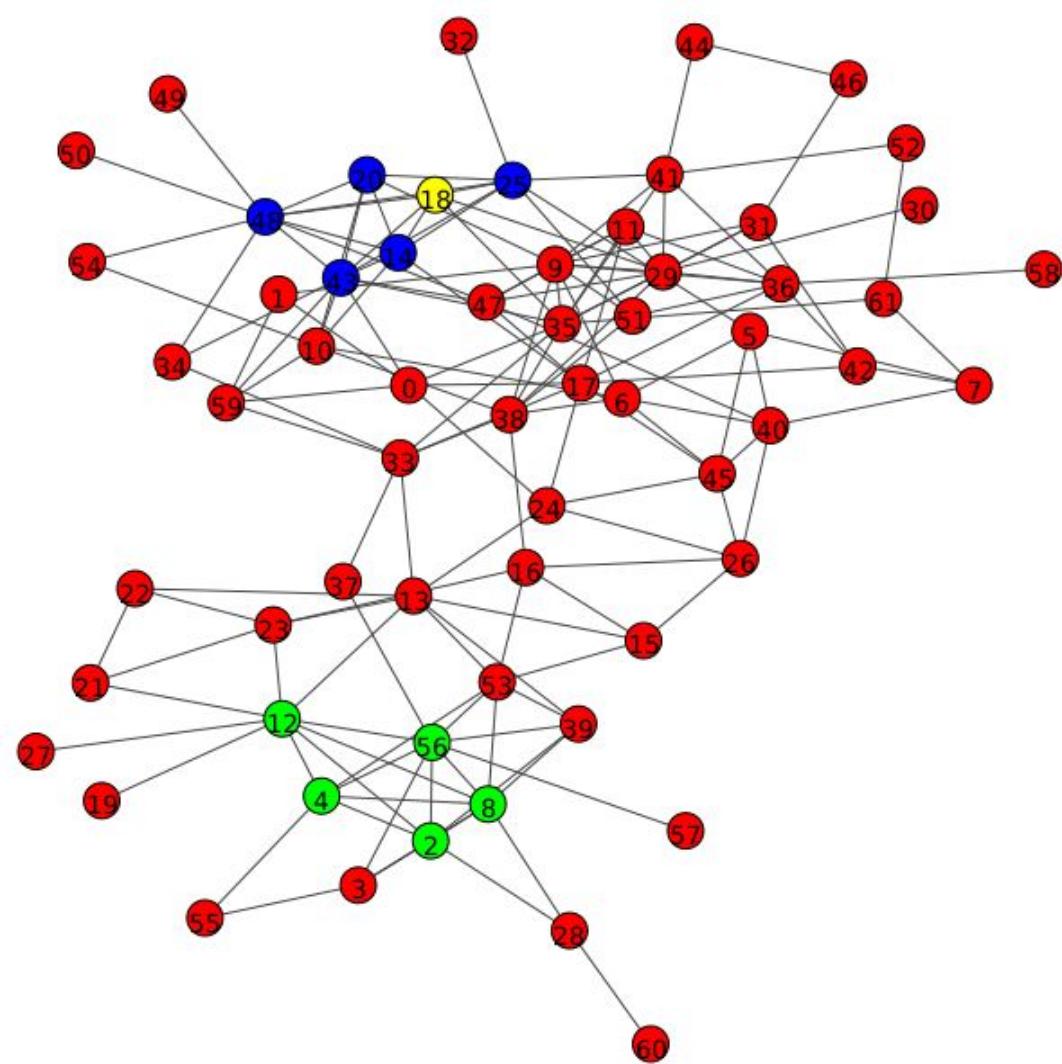
DataSet : Dolphins

Graph Property : **Core**

Given a graph G = (V,E) a subset of nodes U is said to be an N-core iff minimum degree in U is at least N.

Gd.k_core()

returned : list of graphs with k cores, k=1,2….n.

This is a 5-Core plot where subset of nodes have minimum degree of at least 5.

We can see that all the cliques of size 5 were included.
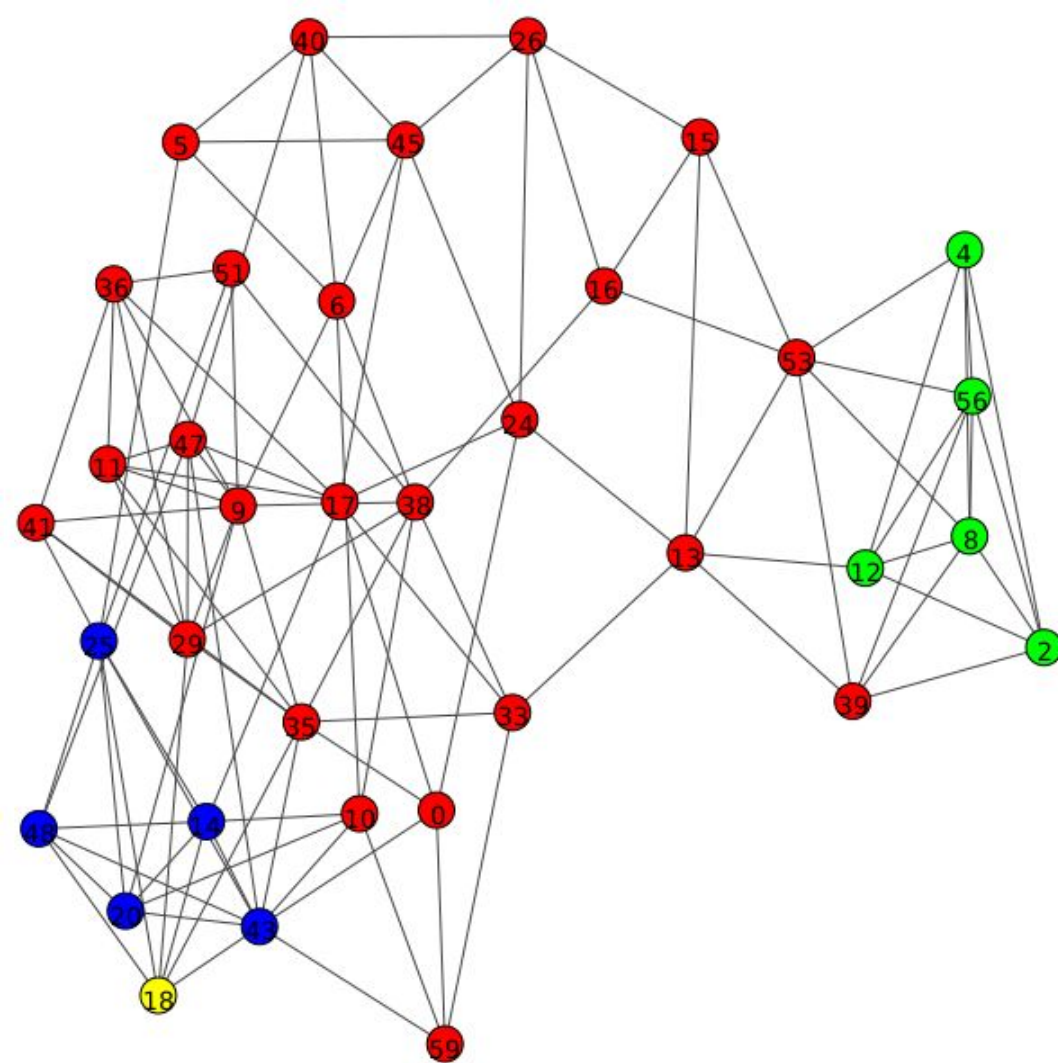
DataSet : Dolphins

Graph Property : **Plex**

A "plex" is a generalization on the concept of a clique, where nodes are allowed to "miss" having edges with other nodes.

Also any N-core is an (n-N)-plex.

1-Plex is also called as Clique Since in a clique for n nodes there should be (n-1) connections. So we are allowed to miss 1.
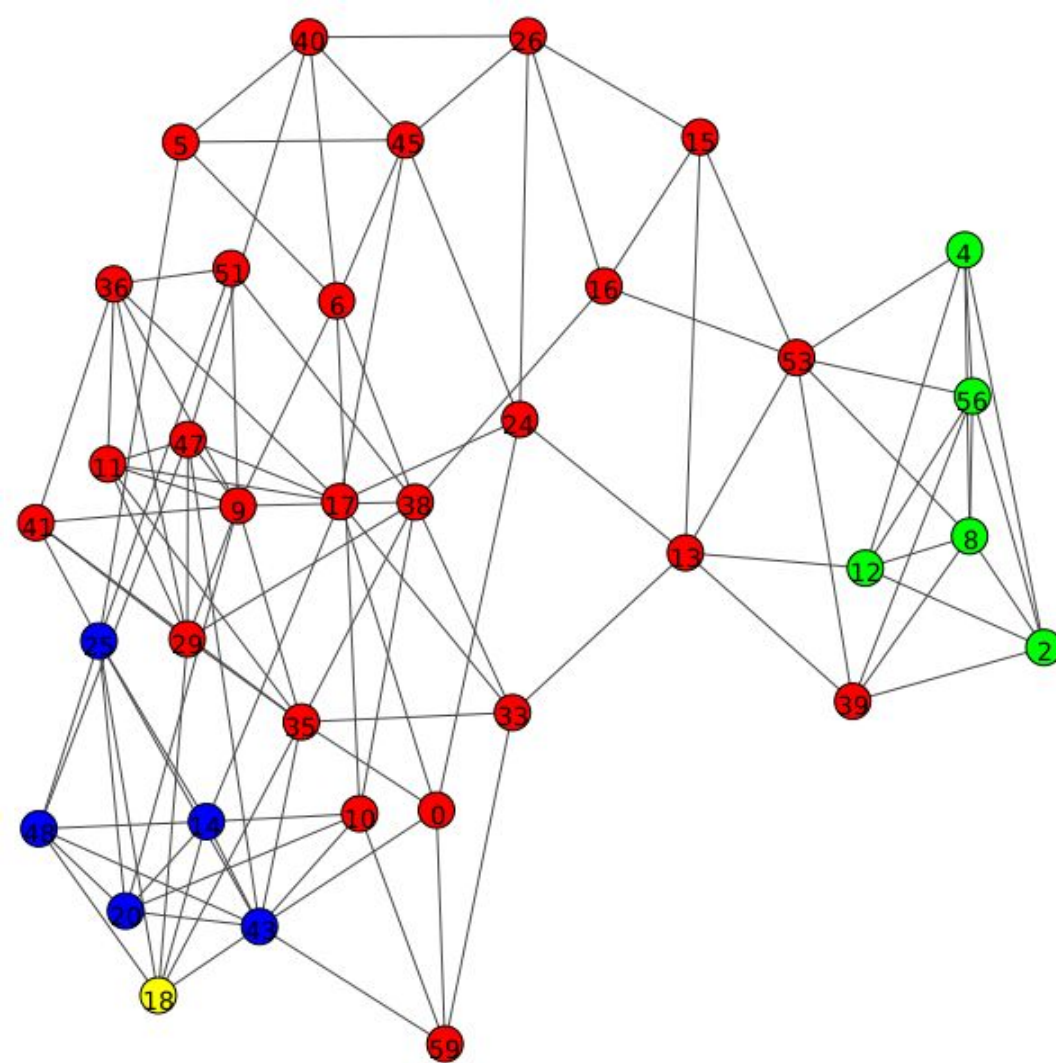
DataSet : Highschool

This directed network contains friendships between boys in a small highschool.

A node represents a boy and an edge between two boys shows that the left boy chose the right boy as a friend. The edge weights show how often that happened.

Max Clique set : [(55, 6, 62, 5, 26, 61),
(27, 45, 66, 44, 67, 65),
(27, 45, 66, 44, 67, 42)]

We see there is a direct edge 36 -> 27 and 36 -> 55. Where 36 has max Betweenness centrality.

So max clique around the node of max centrality is actually a max clique.
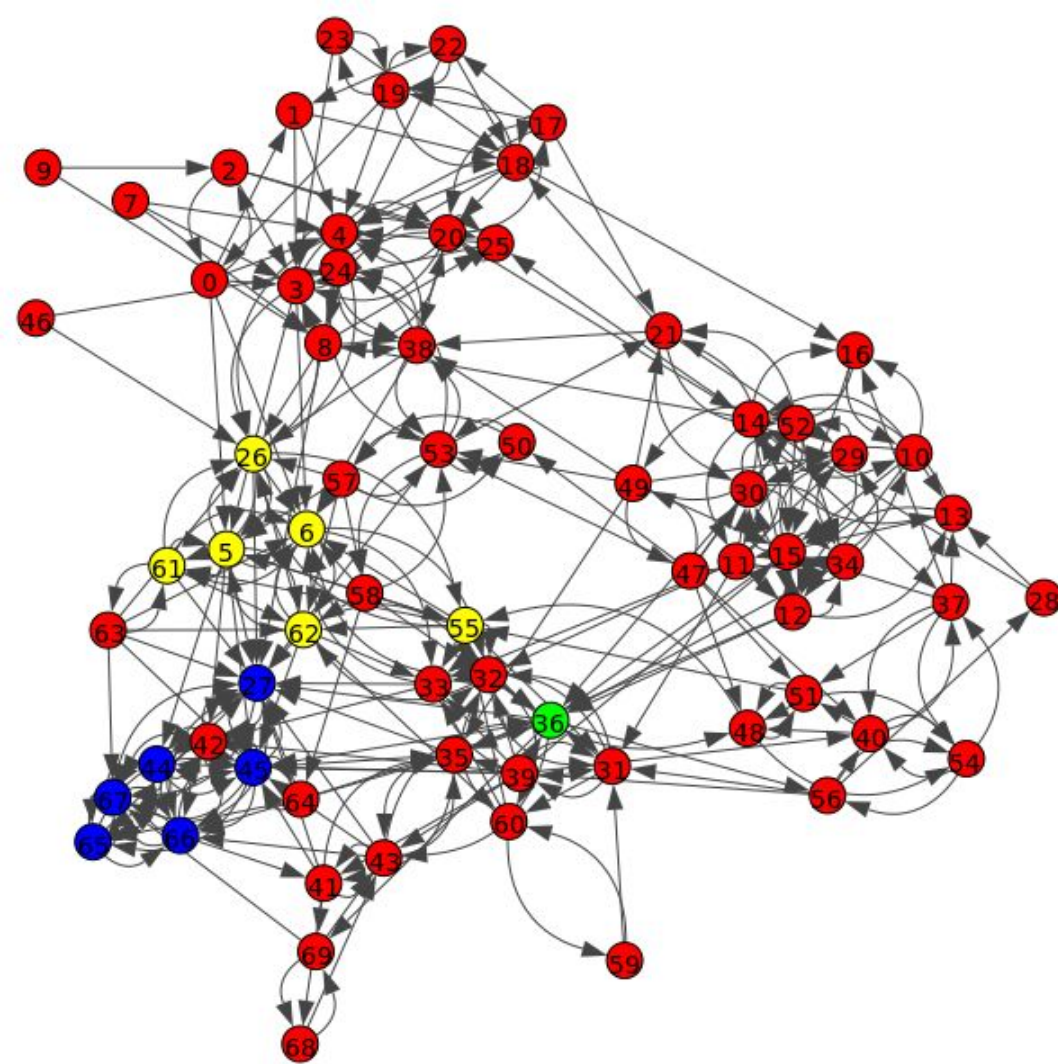
DataSet : Dolphins

Some **interesting insights**.

Adult males live mostly alone or in groups of **two to three**, and join pods for short periods of time.

Adult females and young dolphins normally live in groups of up to **15** animals.

However, they **live in fission-fusion societies** of varying group size, within which individuals change associations often on a daily or hourly basis.

Females with their recent offspring, older subadults of both sexes and adult males **either alone or in bonded pairs.**
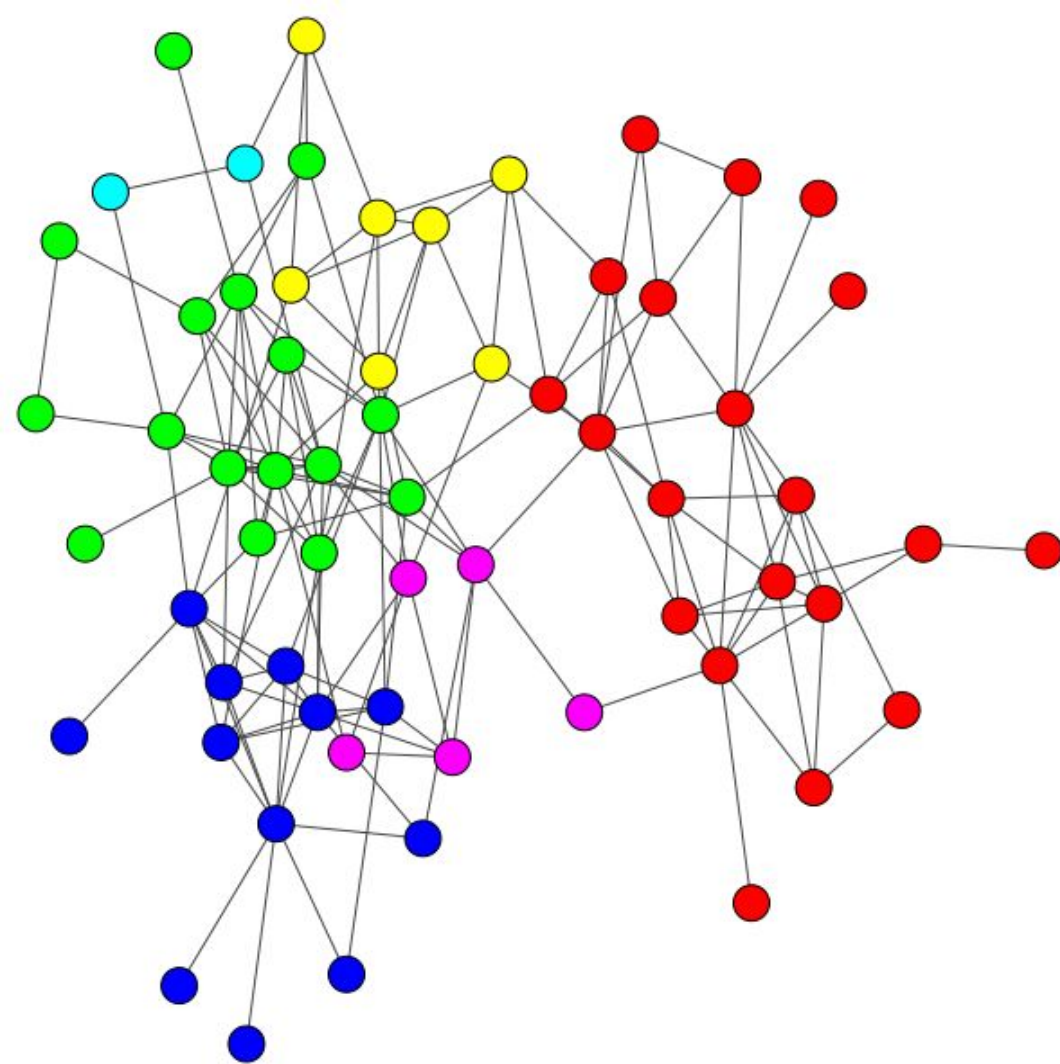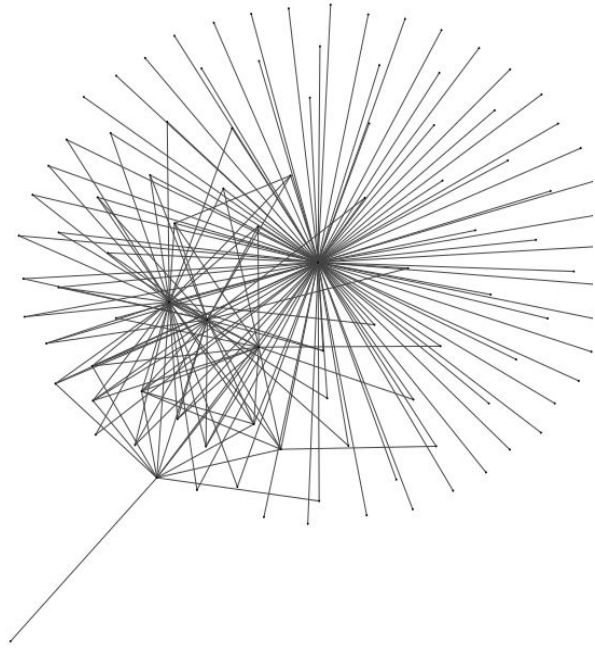
100 Nodes
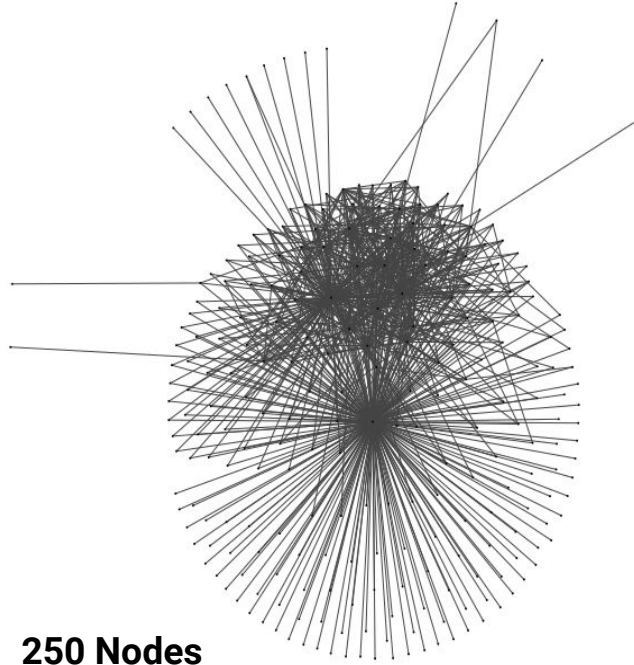
250 Nodes
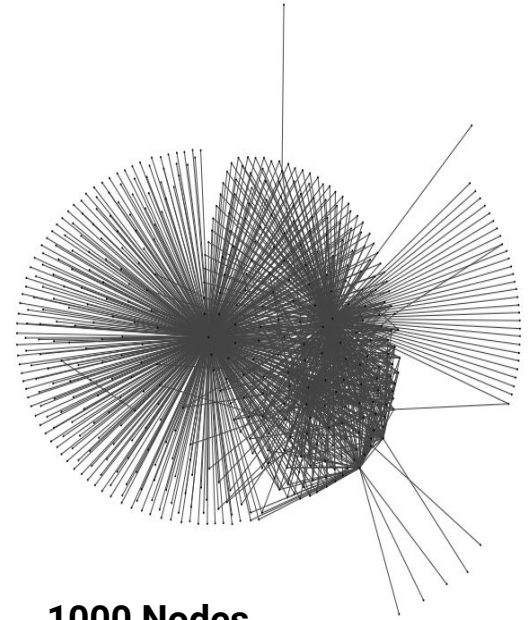
1000 Nodes

Graph with power law degree distribution - scale free at different levels.

Moreno Highschool

This directed network contains friendships between boys in a small highschool in Illinois. Each boy was asked once in the fall of 1957 and the spring of 1958. This dataset aggregates the results from both dates. A node represents a boy and an edge between two boys shows that the left boy chose the right boy as a friend. The edge weights show how often that happened. As a boy could choose the same boy twice edge values from 1 to 2 are allowed.
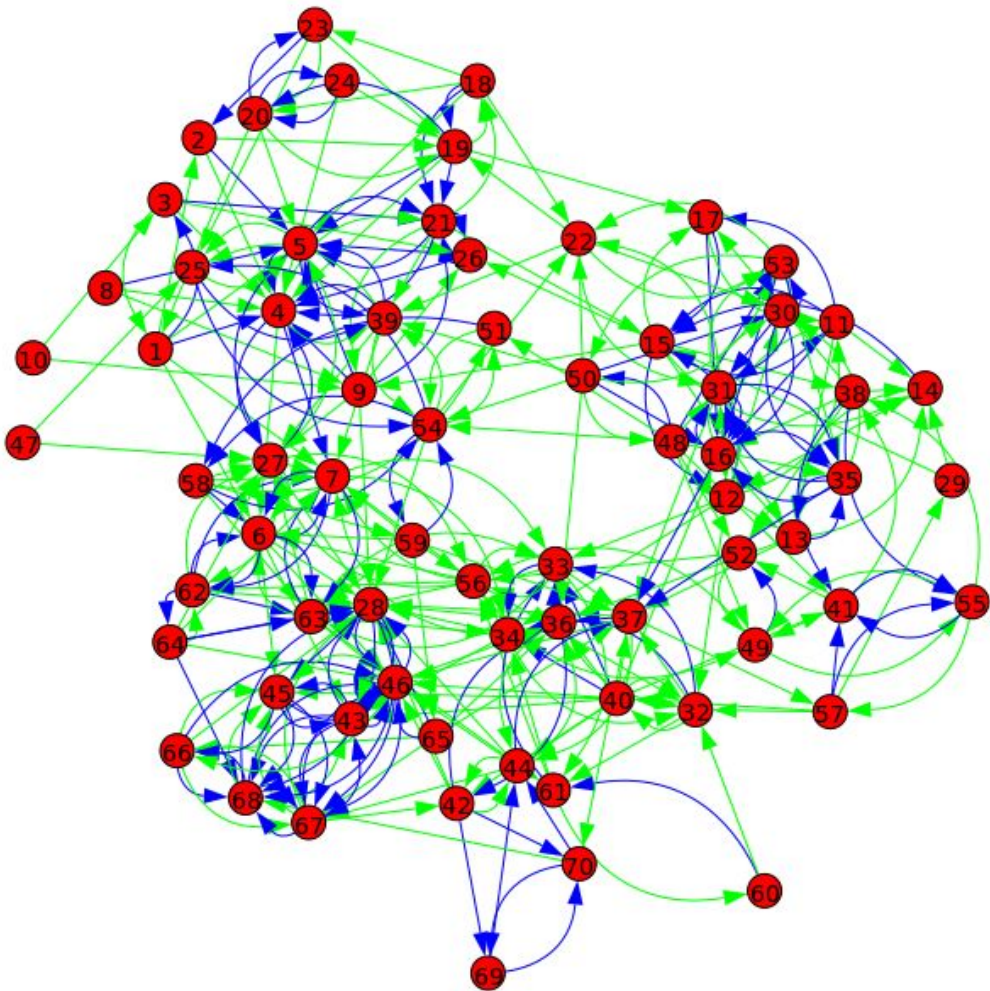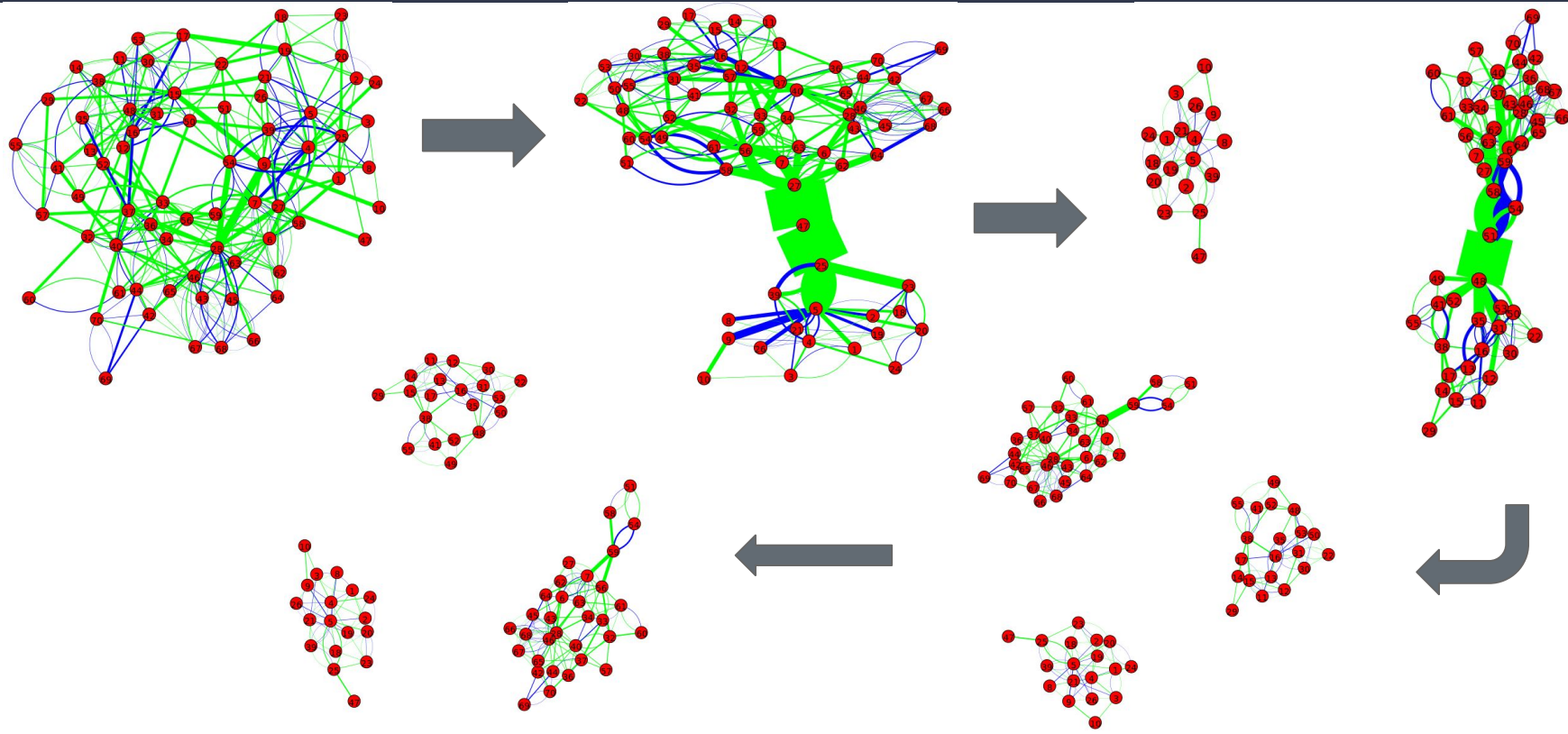
Assumptions

The first survey was done shortly after the batch was formed
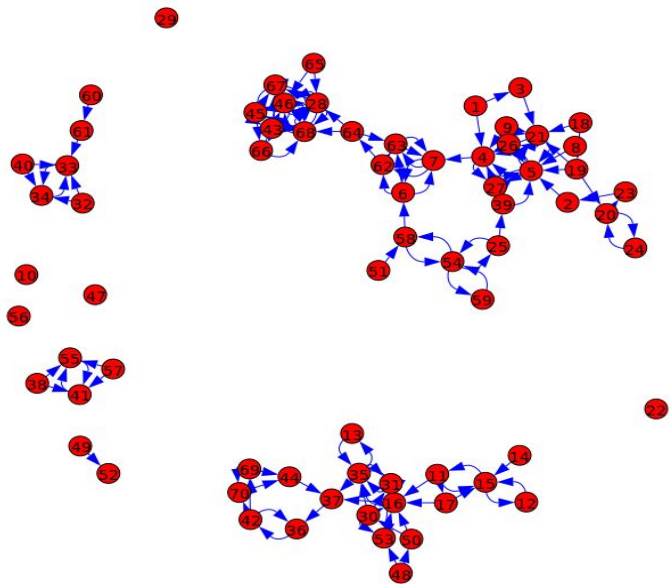
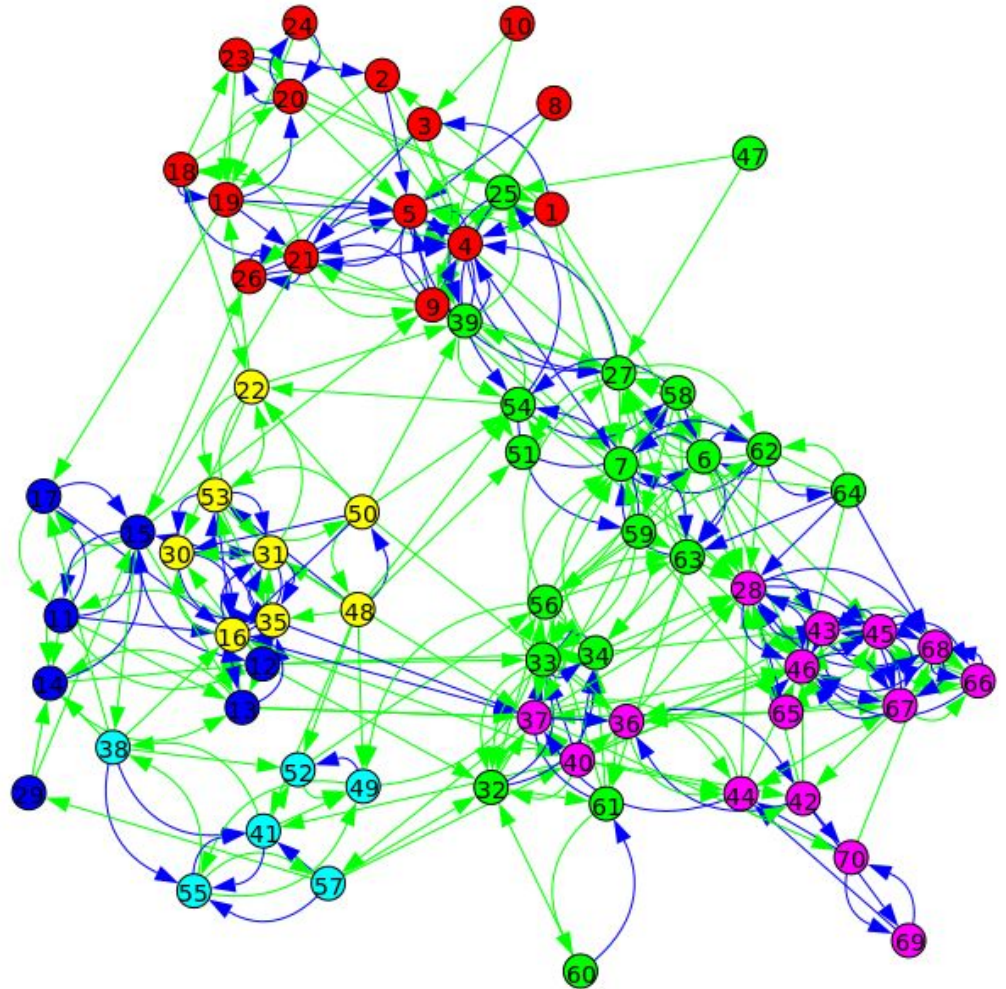**Most of the edges with number weight 1 are from the new connections that are formed.**
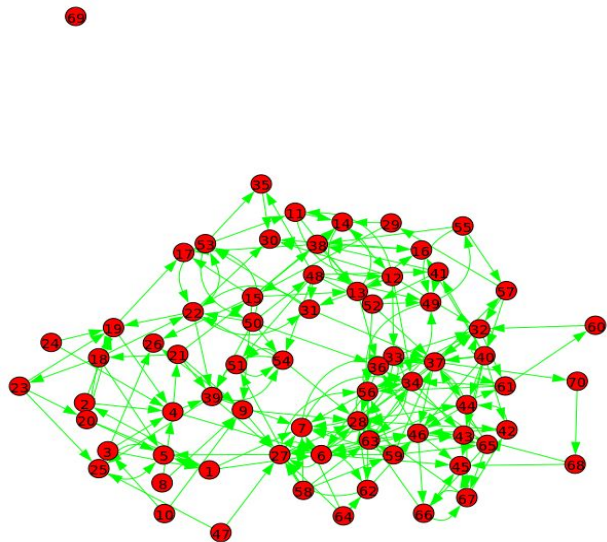
Community detection

Community clustering

It can be seen that most of the blue edges are intra-community which indicates the relations formed early on determine the communities.
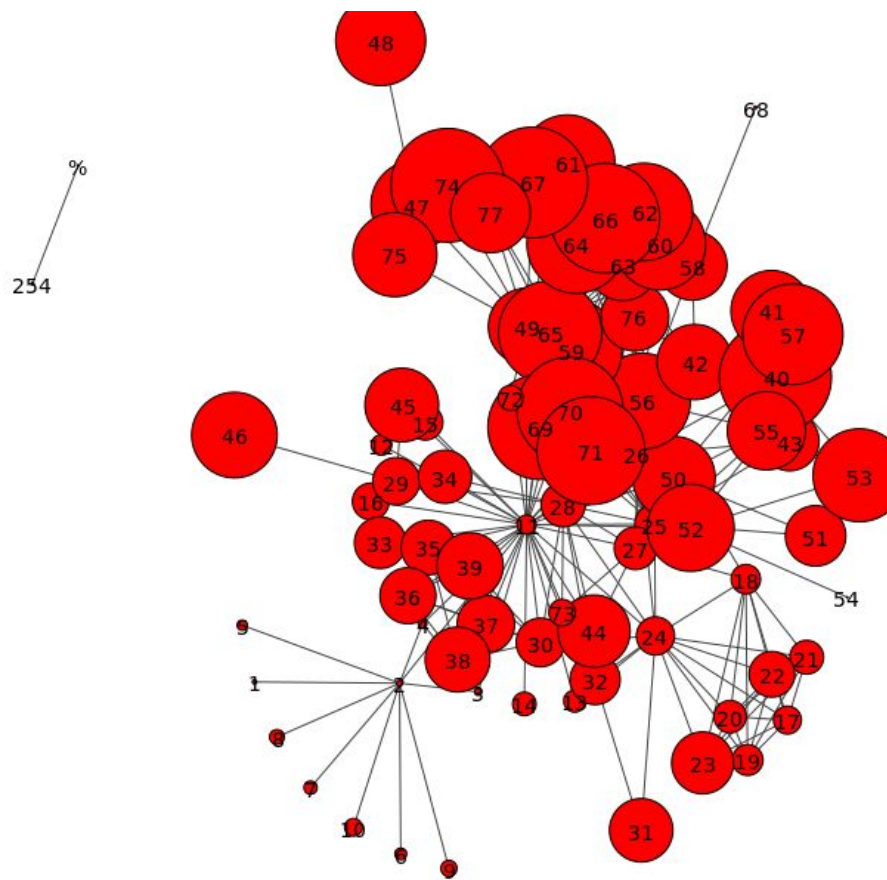
DATASET :   'Les Misérables'

This undirected network contains co-occurances of characters in Victor Hugo's novel 'Les Misérables'.
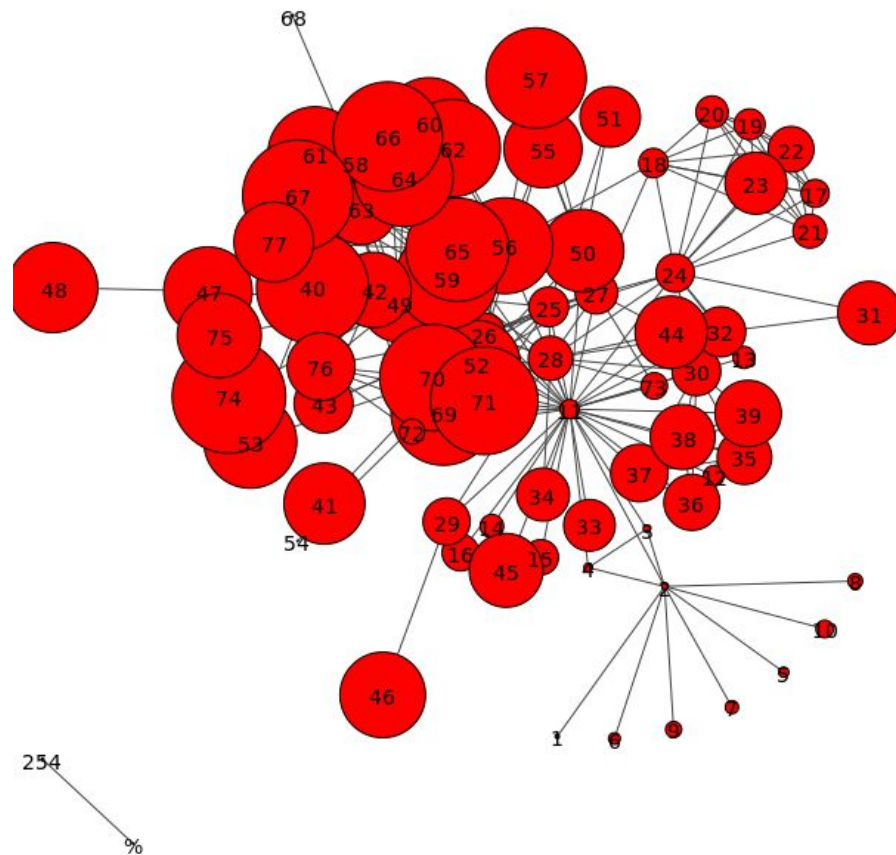ai

A node represents a character .(77 nodes)

An edge between two nodes shows that these two characters appeared in the same chapter of the the book. (254 edges)

The weight of each link indicates how often such a co-appearance occured.

betweenness centrality

betweenness centrality(random walk)

a node with higher **betweenness centrality** would have more control over the network, because more information will pass through that node.

closeness centrality

closeness centrality(random walk)

Way of detecting nodes that are able to spread information very efficiently through a graph.
Nodes with a high closeness score have the shortest distances to all other nodes.

Contiguous usa dataset

48 vertices(states)

107 edges(borders)

Edge betweenness:thickest edge,more edge betweenness,more shortest paths
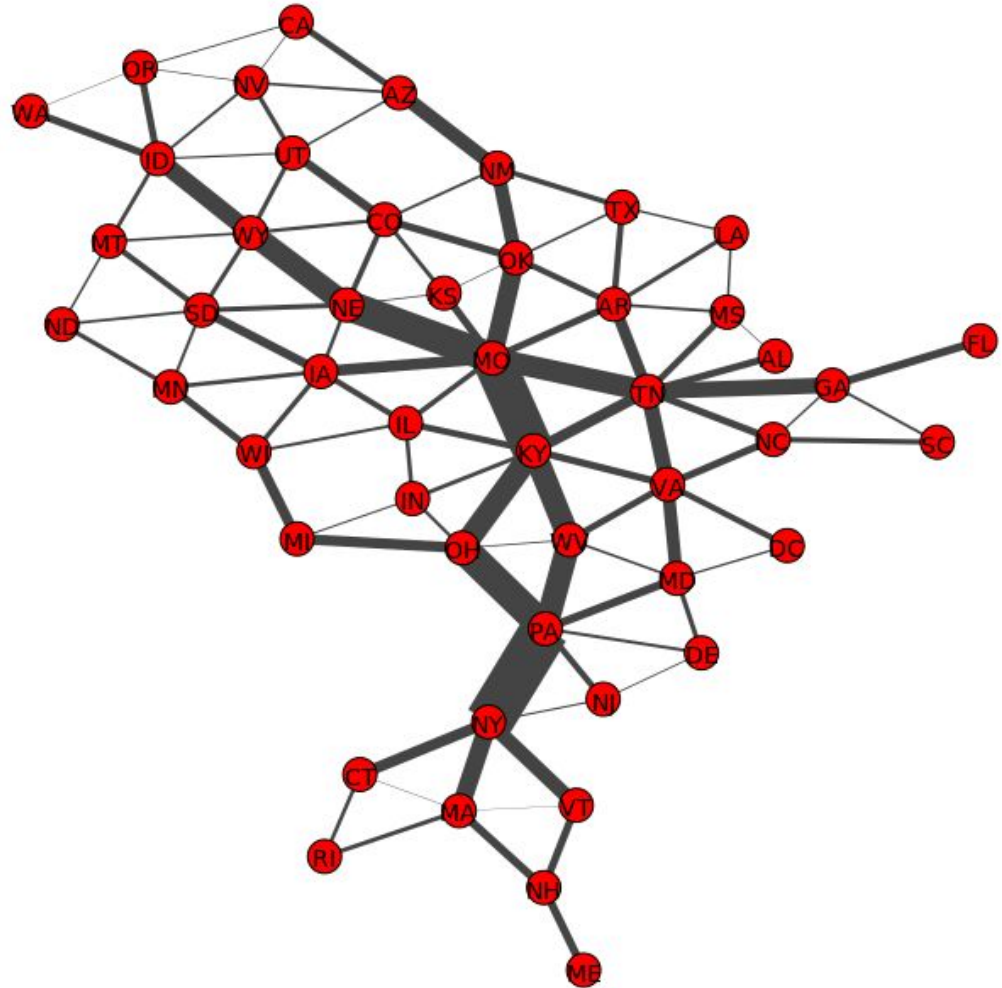
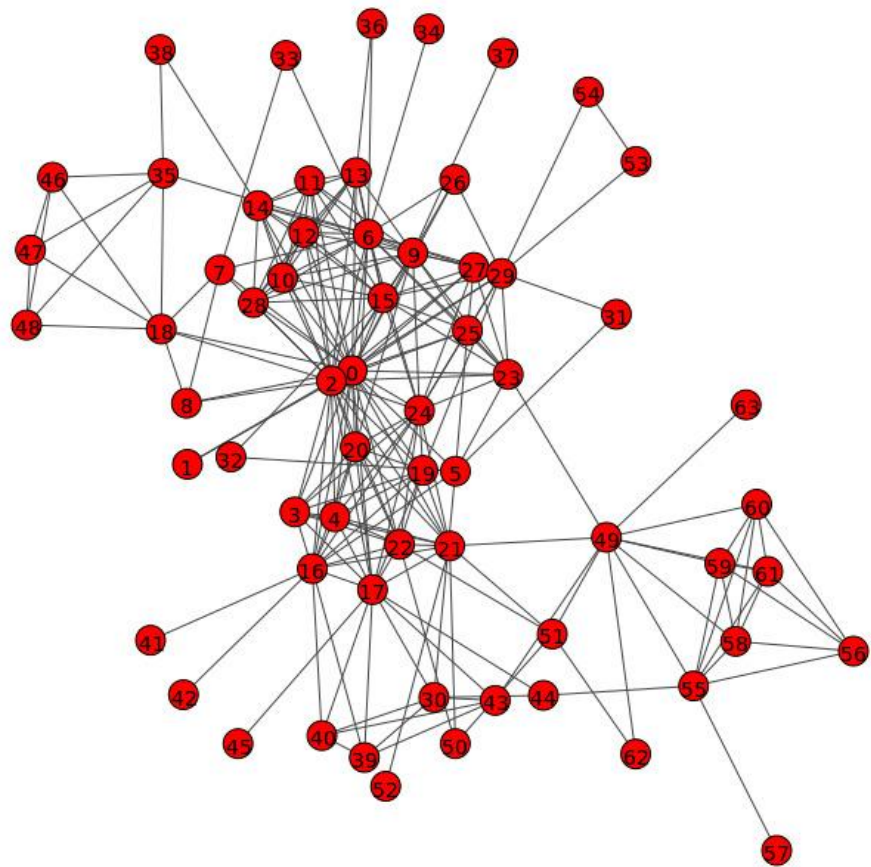Wyoming   21.2 deaths per 100,000 popl
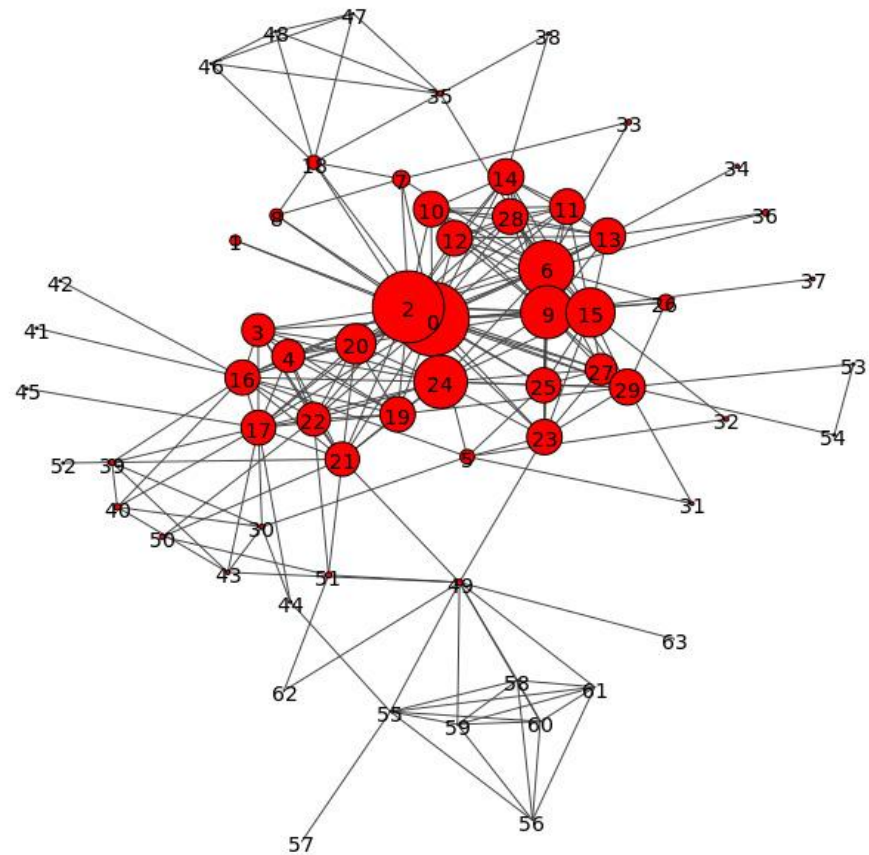
Montana 17.7

Kentucky 17.6

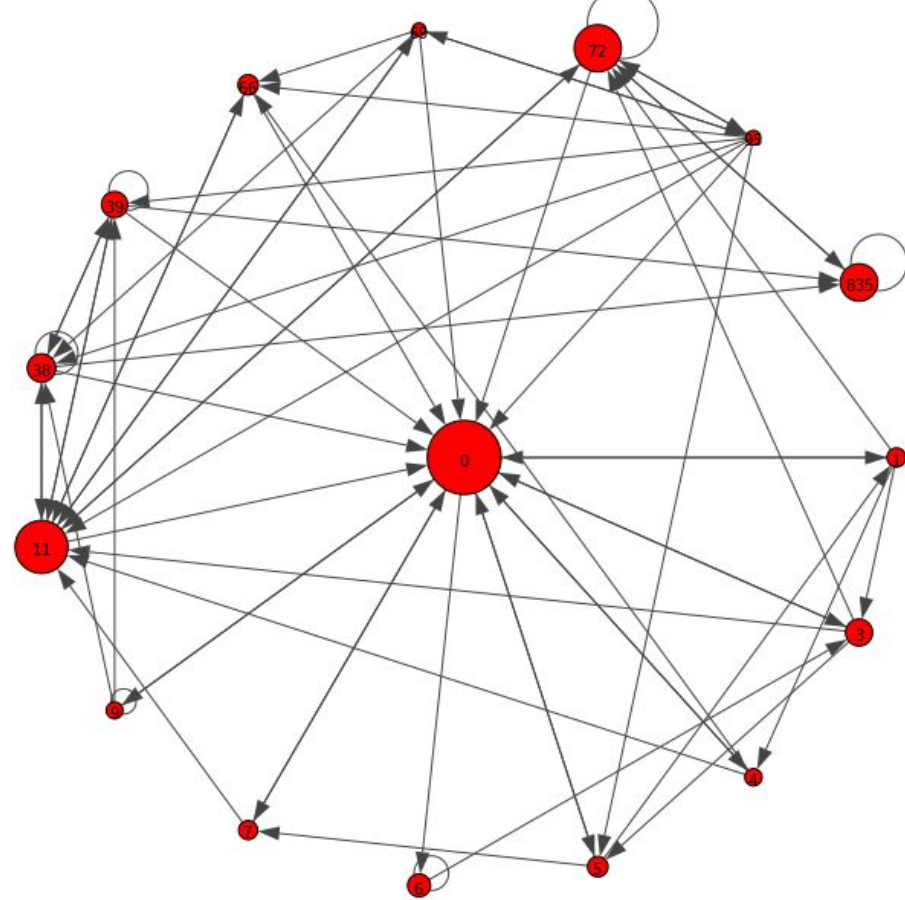Oklahoma 16.7

Tennesse 15.5

Nebraska 11.9

Degree centrality

Bonacich centrality

CFinder-Google Dataset

Pagerank: It evaluates the quality and quantity of links to a webpage to determine a relative score of that page's importance.

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$
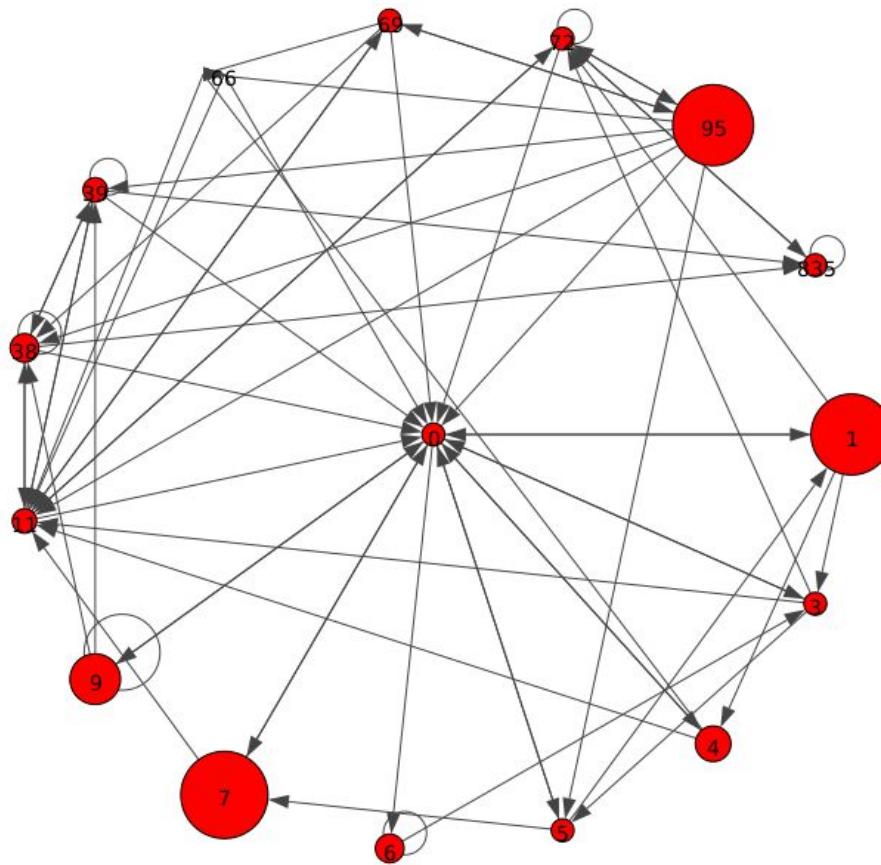

Page rank plot

HITS

HITS(Hypertext-Induced Topic Selection) :It is a measure of importance of pages or documents, similar to PageRank.

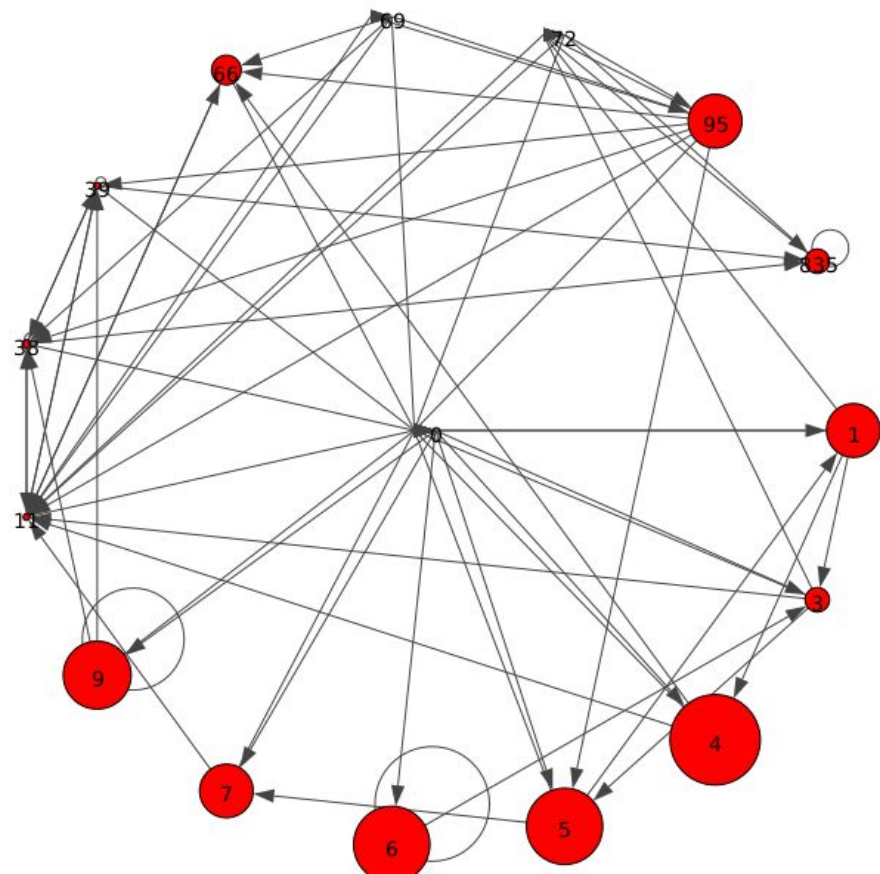**Authorities** are pages containing useful information.

Authority update:
$$\text{auth}(p) = \sum_{q \in P_{to}} \text{hub}(q)$$

Hubs are pages that link to authorities.

Hub update:
$$\text{hub}(p) = \sum_{q \in P_{from}} \text{auth}(q)$$

Hits plot based on Authority value

Hits plot based on Hub value