

**SE 434 01**  
**Social Network Analysis**

**Complete Network Analysis of  
Currently Working American Actors  
& Actresses and Their Filmographies**

Hilal Ayışığ1  
Işıl Deniz Öztürk

**Asst. Prof. Dr. Volkan TUNALI**

## **Abstract**

In this study, we analyzed American actors and actresses who are still acting today based on the data we used. We gathered data which are actors and actresses and their filmography from Ranker and IMDB websites manually.

Then we wrote a python code for constructing the network based on these data and modified the code in order to visualize and analyze our network by a Network Visualization tool. We used two files which are the nodes and edges as CSV format, to import into Gephi for visualization.

In many movies, we can see that some actors/actresses act together more than once. Our aim was to determine which American actors/actresses we used as our data are acted together in the same movie, and by using this information we wanted to show the relationship between actors/actresses based on the movies that act together.

## **Introduction**

The cinema is an effective tool for the audience. A great success has been achieved in spreading the dominant ideology to the masses. The film and video industry has been an increasingly strenuous sector, with its ever-expanding range, entertainment sector, and the growing TV series industry supporting the film industry.

Hollywood movies are the first to come to mind when it comes to motion pictures, the cinema.

Hollywood which is commonly referred for the cinema in the United States, is recognized as the world's oldest film industry and the largest film industry in terms of revenue. It has had a large effect on the film industry in general, and has had a profound effect on cinema across the world since the early 20th century.

European filmmakers from Italy, France, England, or elsewhere have individually competed with Hollywood filmmakers for prominence, but by the end of the first century of film, Hollywood commonly absorbed filmmakers and filmmaking from throughout the world within its complex system of producing, financing, and distributing entertainment products. British film producer David Puttnam noted, "Americans understand, better than anyone else, how to produce and market films, just as the Germans make cars, the French make perfume and the Scots make whisky."

Hollywood's movie studios are the primary source of the most commercially successful and best-selling movies in the world.

Today, American film studios produce several hundred films each year, making the United States one of the most prolific filmmakers in the world and a leading leader in motion picture engineering and technology.

Hollywood has gained a great place in the film industry and is recognized worldwide, as a result of this success, American films and their actors are known by many people around the world.

We watch world-famous actors in many American movies and we can notice that the same actors play together again in some movies. In this study, we evaluated the bond between these actors over the films they played together. Considering the films in which world-famous actors and actresses starred together, we decided to create a network with only American actors and their movies. There have been plenty of great American actors and actresses throughout film history. We filtered the data required for this network as American actors and actresses who are actively acting in movies and their filmography, using the data we obtained according to this filtering, we created networks of actors and actresses and analyzed these networks using network analysis tools. In this study, Gephi was used for social network analysis and visualization studies.

## **Data Gathering**

The data used in this study obtained from the article prepared by the ranker site and Imdb site. Despite the filtering we made there are still many American actors, because of that we decided to use only actors and actresses who are still working today for this study. We collected the data of actors and actresses from the best american actors and actresses working today articles prepared by the ranker site, and the filmographies of these actors and actresses from the Imdb site manually.

<https://www.ranker.com/list/best-american-actresses-working-today/ranker-film>

<https://www.ranker.com/list/best-american-actors-working-today/ranker-film>

[https://www.imdb.com/?ref\\_=nv\\_home](https://www.imdb.com/?ref_=nv_home)

## Network Construction

We defined a list called Actors and inside of this we kept separate lists containing the movie names of each actor/actresses. After defining this list that contained movies of actors, we defined another list that stored the names of the actors and actresses according to the order of the movie lists in this actor list. Later, we wrote a python code to find the actors who starred in a movie together and defined a dictionary to store the actors and movies that they starred together in this dictionary. In the network we have created, the nodes represent the actors and actresses and the edges represent the movies they have starred together, we assumed the actors and actresses who took part in the same movie have a connection with this evaluation, we combined the nodes of these actors with the edges, and we used the number of movies they played together as the weight of the edges between these nodes.

Sample of the lists that we created:

```
actors = [  
    ["Lucy in the Sky", "Avengers: Endgame", "The Death & Life of John F. Donovan", "Vox, ...]  
  
    ,["The Witches", "The Last Thing He Wanted", "Dark Waters", "Modern Love", ... ]  
  
    ,["Jojo Rabbit", "Marriage Story", "Avengers: Endgame", "Captain Marvel" ...]  
]  
  
actorid = ['Natalie_Portman', 'Anne_Hathaway', 'Scarlett_Johansson', .....]
```

The network consisted of the nodes which represented the actors and actresses, and the edges which represented the relationships between these actors and actresses based on the movies they have starred together. We wrote python code and by importing the csv library we were able to convert our nodes and edges list to csv format in order to meet the importing data specification of the network visualization tool called Gephi. The Gephi program allows us to import these lists in csv format easily.

Id	Label
1	Natalie Portman
2	Anne Hathaway
3	Scarlett Johansson
4	Charlize Theron
5	Emma Stone
6	Sandra Bullock
7	Meryl Streep
8	Reese Witherspoon
9	Jennifer Lawrence
10	Betty White
11	Julia Roberts
12	Michelle Pfeiffer
13	Julianne Moore
14	Jessica Chastain
15	Diane Lane
16	Amanda Seyfried
17	Jennifer Aniston
18	Kristen Bell
19	Jennifer Garner
20	Mila Kunis
21	Elizabeth Banks
22	Amy Adams
23	Jessica Lange
24	Viola Davis
25	Zoe Saldana

**Figure 1:** Node.csv file format sample

Id	Source	Target	Type	Weight
1	Scarlett Johansson	Natalie Portman	Undirected	2.0
2	Emma Stone	Scarlett Johansson	Undirected	1.0
3	Sandra Bullock	Anne Hathaway	Undirected	1.0
4	Meryl Streep	Anne Hathaway	Undirected	1.0
5	Reese Witherspoon	Meryl Streep	Undirected	3.0
6	Jennifer Lawrence	Anne Hathaway	Undirected	1.0
7	Jennifer Lawrence	Charlize Theron	Undirected	1.0
8	Betty White	Sandra Bullock	Undirected	1.0
9	Betty White	Meryl Streep	Undirected	1.0
10	Julia Roberts	Meryl Streep	Undirected	3.0
11	Julia Roberts	Anne Hathaway	Undirected	1.0
12	Julia Roberts	Natalie Portman	Undirected	2.0
13	Julia Roberts	Reese Witherspoon	Undirected	1.0
14	Michelle Pfeiffer	Natalie Portman	Undirected	1.0
15	Michelle Pfeiffer	Scarlett Johansson	Undirected	1.0
16	Michelle Pfeiffer	Jennifer Lawrence	Undirected	1.0
17	Michelle Pfeiffer	Betty White	Undirected	1.0
18	Michelle Pfeiffer	Sandra Bullock	Undirected	1.0
19	Julianne Moore	Jennifer Lawrence	Undirected	2.0
20	Julianne Moore	Scarlett Johansson	Undirected	1.0
21	Julianne Moore	Emma Stone	Undirected	1.0
22	Julianne Moore	Meryl Streep	Undirected	1.0
23	Jessica Chastain	Jennifer Lawrence	Undirected	1.0
24	Jessica Chastain	Charlize Theron	Undirected	1.0
25	Jessica Chastain	Anne Hathaway	Undirected	1.0

**Figure 2:** Edge.csv file format sample

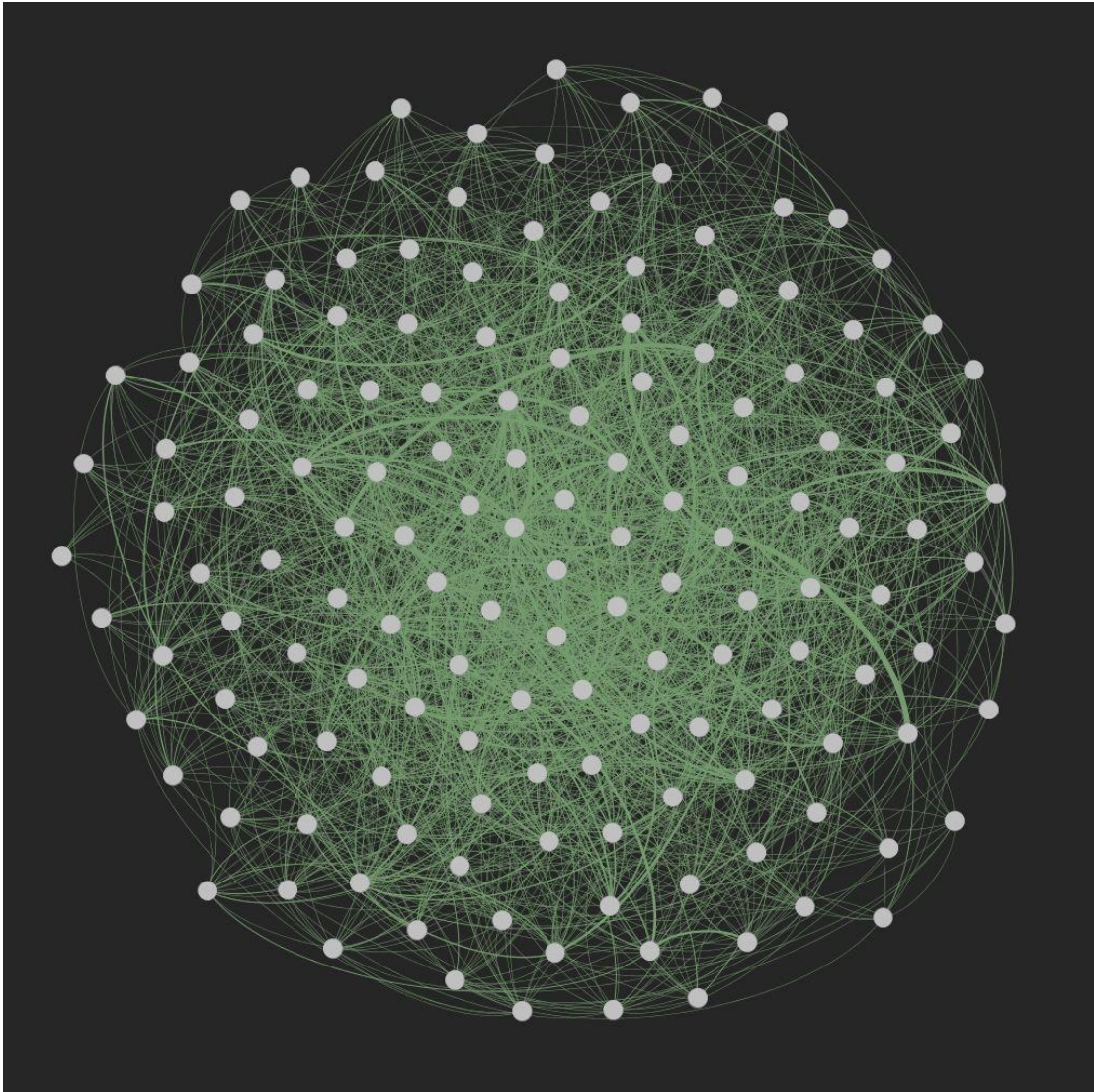
Sample of the edge list formed according to the relationships we have determined between the actors/actresses using the python code we wrote, and the sample node list of the 150 actors/actresses we have collected is shown. We created our network as undirected because our network is formed by combining the nodes of the actors, who are starred together in some movie, with the edge.

## Network Analysis

Network analysis is a set of integrated techniques to depict relations among actors and to analyze the social structures that emerge from the recurrence of these relations. In this study, the network analysis was done by dividing it into several parts. By doing this analysis we can more easily understand the functionality of data in the network.

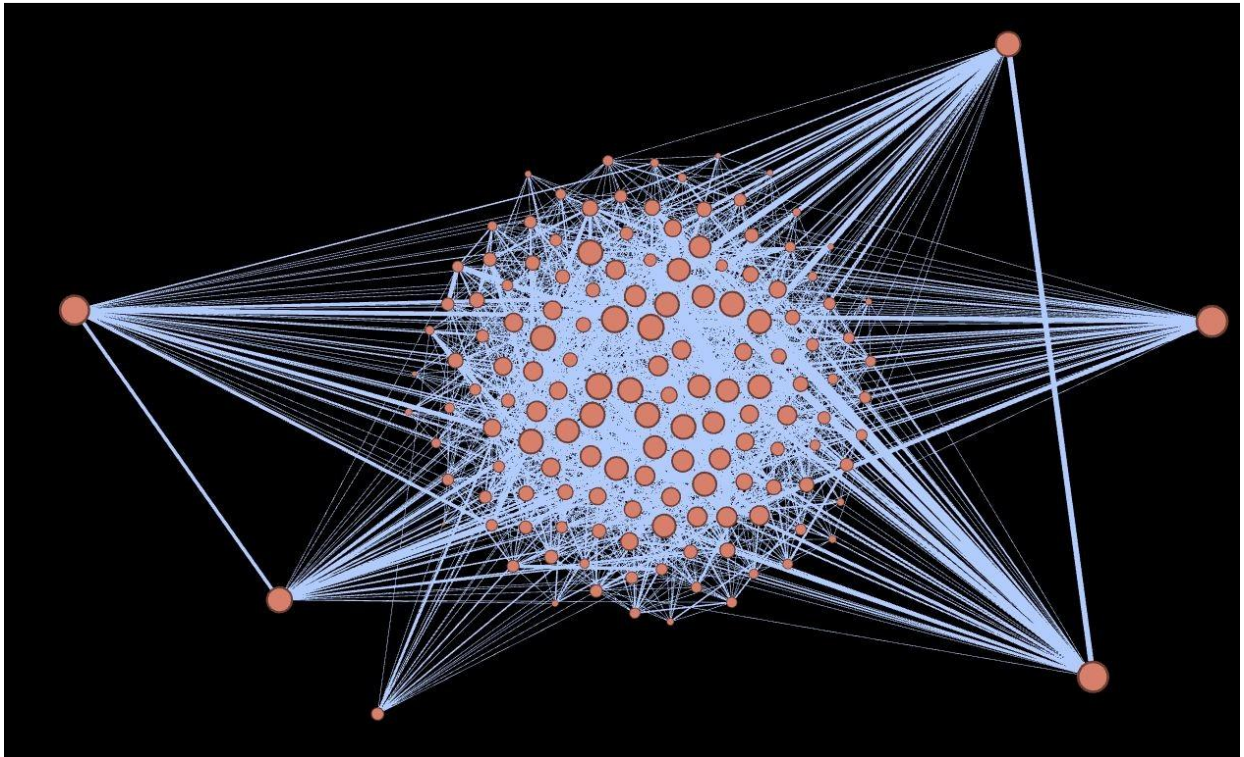
## 1. Visual Analysis

We employ a standard force-directed algorithm as a basis, that is, the algorithm of Fruchterman and Reingold. Because of our network being undirected weight between nodes  $x \leftrightarrow y$  is the same.



**Figure 3:** *Visualization of the Network of Currently Working American Actors & Actresses and Their Filmographies*





**Figure 4:** *Visualization of the Weighted Network of Currently Working American Actors & Actresses and Their Filmographies*

The edges are associated with weights. In this study the weight associated with each edge could represent the number of movies which the two nodes of actors/actresses played together.

## **2. Centrality Analysis**

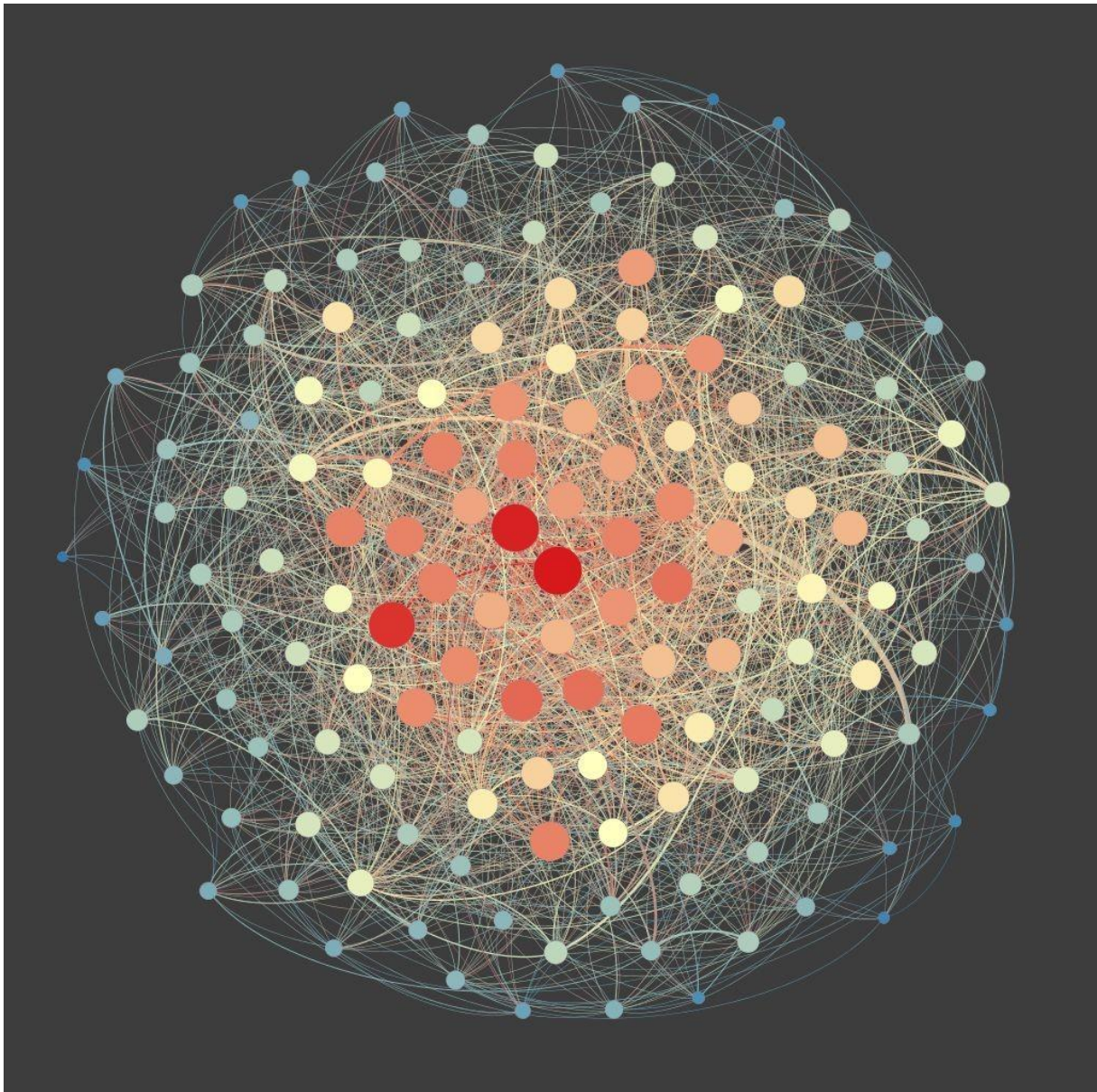
Centrality defines how important a node is in a network. It helps us to calculate the importance of any given node in a network. In this study it determines the most important or prominent actors/actresses(nodes) in the network based on their location. Each measurement has its own definition of importance so we will look at each of these centrality measurements based on our network.

## 2.1 Degree Centrality

Rank	Actor/Actresses	Degree
1	Susan Sarandon	60
2	Samuel L. Jackson	59
3	Stanley Tucci	57
4	Paul Rudd	51
5	Mark Ruffalo	50
6	Christopher Walken	50
7	Woody Harrelson	49
8	Bradley Cooper	48
9	Don Cheadle	48
10	Marisa Tomei	48
11	Chris Pratt	48
12	Robin Wright	48
13	Ben Affleck	48
14	Meryl Streep	48
15	Michelle Pfeiffer	47

**Table 2:** The most important top 15 nodes based on Degree Centrality measurement





**Figure5:** *Visualization of the network based on degree centrality measurements.*

Degree is the number of direct connections, neighbors of a node with this information degree centrality defines the importance of a node. Based on degree values the table shows us the most important 15 actors/actresses in the network.

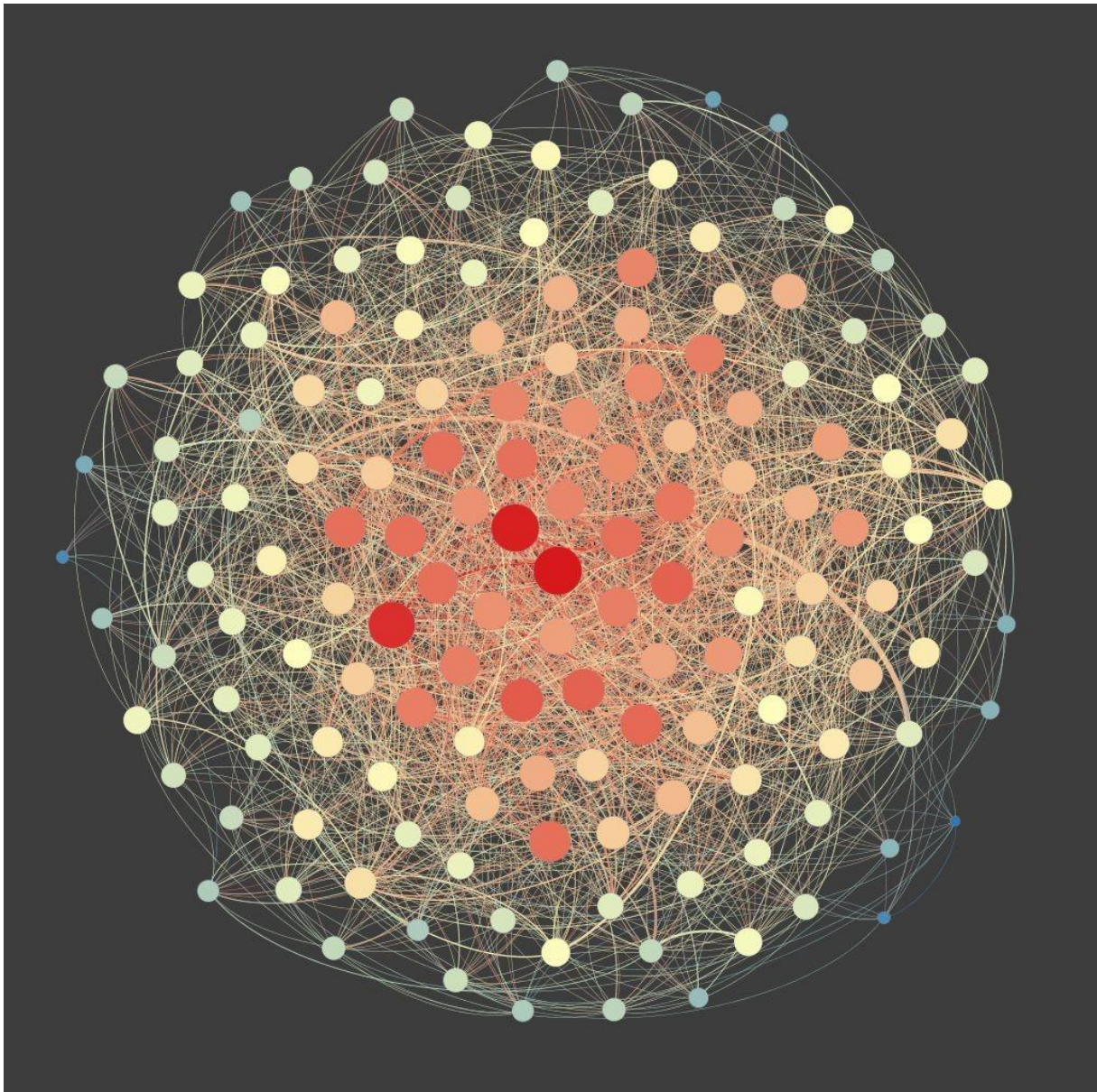
The size and color of the nodes change according to the number of degrees the nodes have. As you go from blue to red, from small node to big node as well, the number of degrees the nodes has increases.

## 2.2 Closeness Centrality

<b>Rank</b>	<b>Actor/Actresses</b>	<b>Closeness</b>
1	Susan Sarandon	0.62605
2	Samuel L. Jackson	0.623431
3	Stanley Tucci	0.618257
4	Paul Rudd	0.603239
5	Mark Ruffalo	0.600806
6	Christopher Walken	0.600806
7	Woody Harrelson	0.598394
8	Robin Wright	0.596
9	Meryl Streep	0.596
10	Marisa Tomei	0.596
11	Don Cheadle	0.596
12	Chris Pratt	0.596
13	Bradley Cooper	0.596
14	Ben Affleck	0.596
15	Robert De Niro	0.593625

***Table 3: The most important top 15 nodes based on Closeness Centrality measurement***





**Figure 6:** Visualization of the network based on closeness centrality measurements.

Closeness defines how important a node is based on the nodes distance to every other node in the network. If a node is important then its position should be important as well in terms of connecting other nodes. We can say that if it is important it has been in such a position that it can reach to other nodes quickly. The size and color of the nodes change according to the number of closeness the nodes have. As you go from blue to red, from small node to big node as well, the number of degrees the nodes has increases.

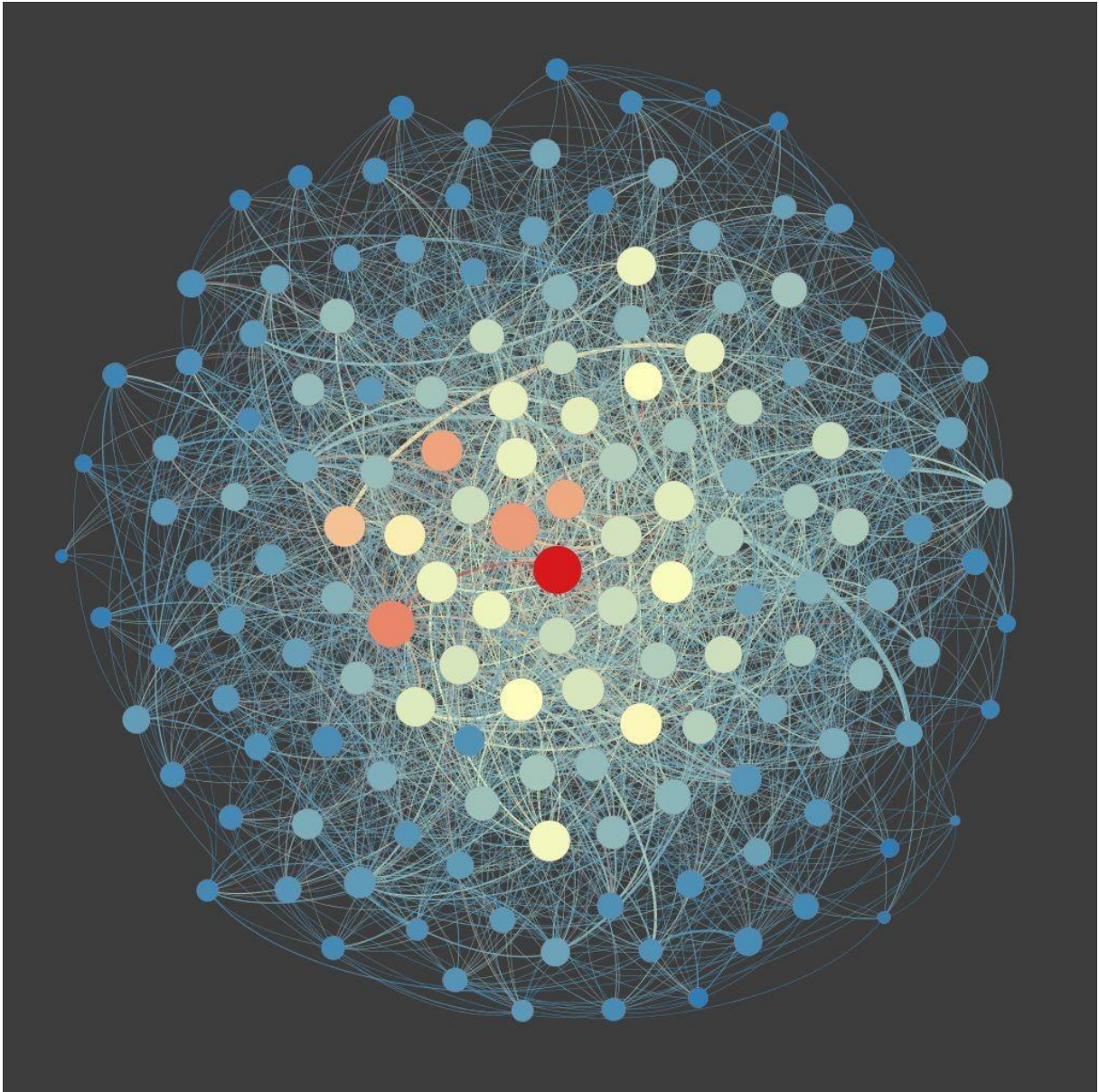
## 2.3 Betweenness Centrality

Rank	Actor/Actresses	Betweenness
1	Susan Sarandon	0.025499
2	Stanley Tucci	0.019457
3	Samuel L. Jackson	0.018263
4	Meryl Streep	0.017852
5	Charlize Theron	0.017522
6	Ben Affleck	0.016193
7	Robin Wright	0.013724
8	Woody Harrelson	0.013226
9	Paul Rudd	0.012955
10	John Malkovich	0.012685
11	Christopher Walken	0.012583
12	Chris Pratt	0.012213
13	Ed Harris	0.011857
14	Angela Basset	0.01181
15	Bradley Cooper	0.011759

**Table 4:** *The most important top 15 nodes based on Betweenness Centrality measurement*

The table shows us the most important 15 actors/actresses in the network based on betweenness values.





**Figure 7:** *Visualization of the network based on betweenness centrality measurements.*

Betweenness defines the importance of a node based on how much a node is in-between the others. If a node is positionally on the path of the most of the shortest paths then this node is considered as important according to this centrality. The size and color of the nodes change according to the number of betweenness the nodes have. As you go from blue to red, from small node to big node as well, the number of degrees the nodes has increases.

## 2.4 Eigenvector Centrality

# Eigenvector Centrality Report

---

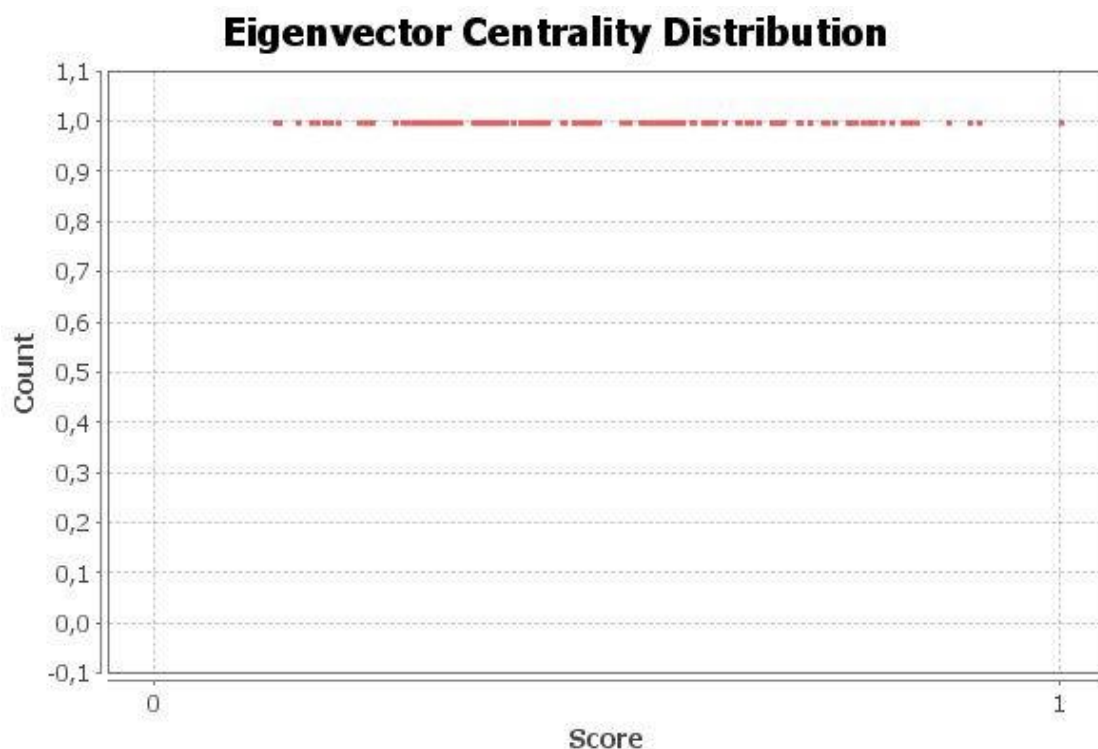
## Parameters:

Network Interpretation: undirected

Number of iterations: 100

Sum change: 0.00224085220329745

## Results:



According to eigen-vector the importance of a node depends not only on the number of neighbors it has, but also on how important its neighbors are. If its neighbors are important as well the number of the neighbors it has then we can say that it is more important.



### 3.Community Analysis

A community is most often defined as a group of individuals living in the same geographical location. It can also be used to describe a group of people with a shared characteristic or common interest. Based on the principle that pairs of nodes are more likely to be connected if they are both members of the same community(ies) and less likely to be connected if they do not share communities.

Community analysis, an activity that involves gathering a wide variety of information about the community.

#### Modularity Report

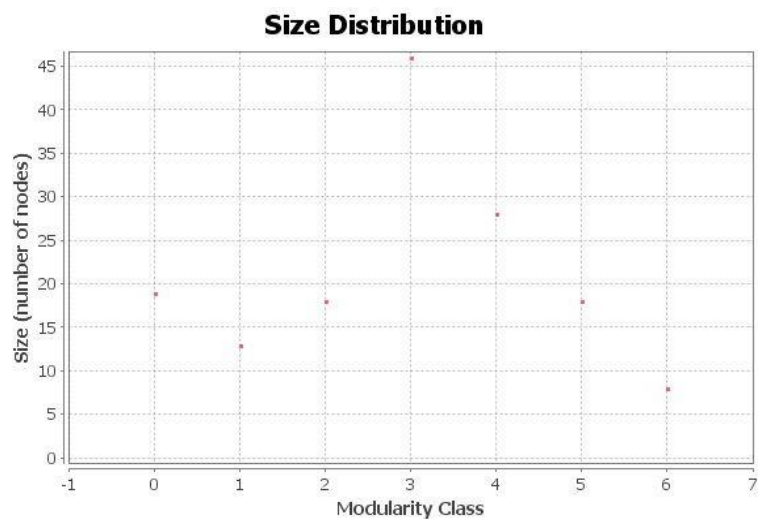
---

##### Parameters:

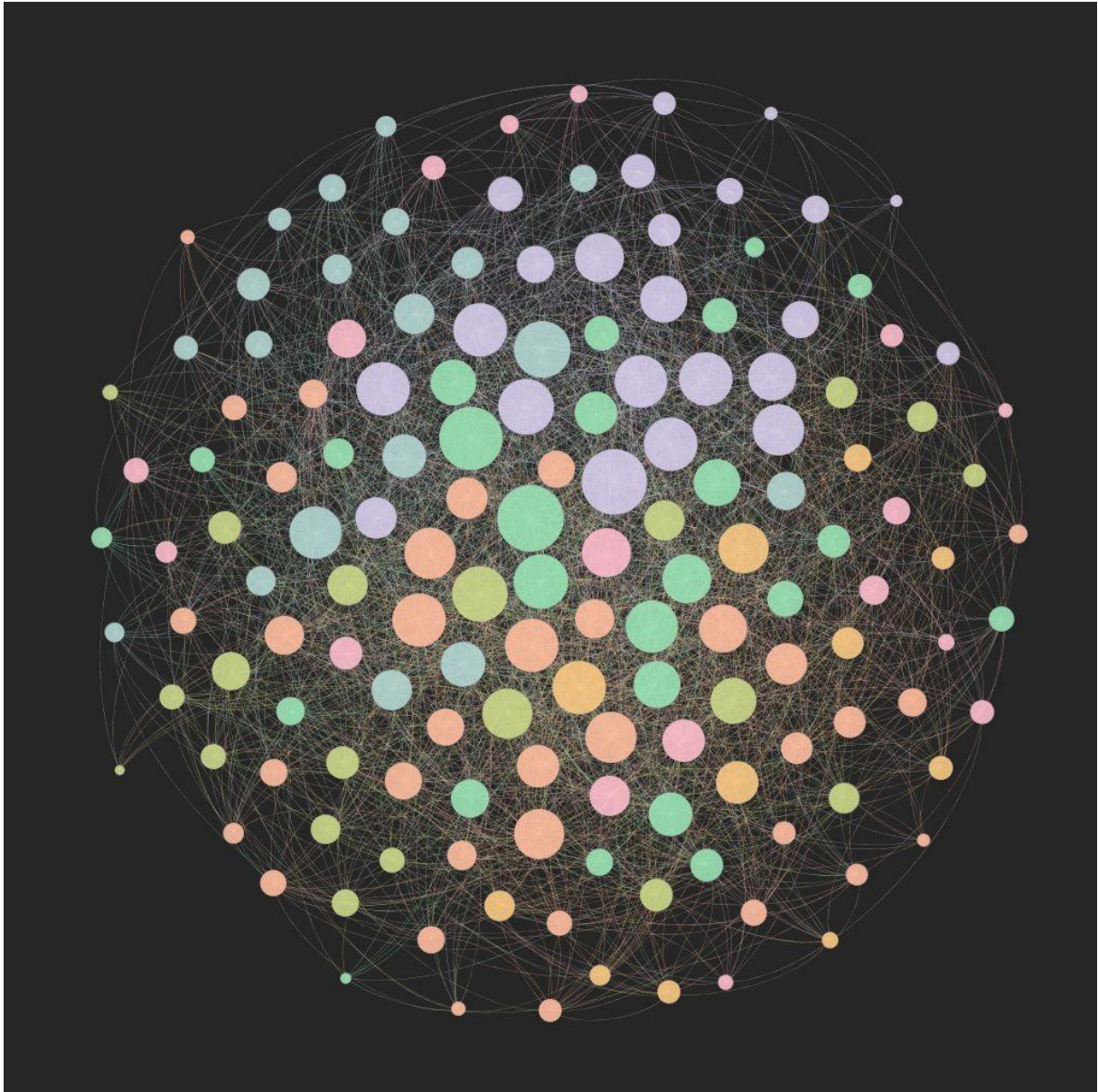
Randomize: On  
Use edge weights: On  
Resolution: 1.0

##### Results:

Modularity: 0.212  
Modularity with resolution: 0.212  
Number of Communities: 7



Modularity measures the strength of division of a network into clusters or communities. Networks with high modularity have dense connections between the nodes within communities.










**Figure 8:** *Visualization of the network based on modularity.*

In the visualization of the network, the sizing of the nodes in proportion to their degrees and also the communities in the network are highlighted with different colors. The node sizes have been selected as min 10 max 70.



**Figure 9:** *A different representation of the modularity based network..*

In addition to the visualization of the network based on the communities in figure 7, this figure is a different representation according to the concept of community.

Community Color	Community ID	Number Of Element
	4	35
	5	25
	0	23
	2	20
	6	19
	3	17
	1	11

**Table 5:** *The table of community colors, ids, and node numbers inside of the community.*

The total number of communities in our network turned out to be 7. Each color is assigned to a different community and to the nodes inside of that community.



## Sample of communities:

### Community 0

Id	Modularity Class	Degree
Samuel L Jackson	0	59
Mark Ruffalo	0	50
Bradley Cooper	0	48
Chris Pratt	0	48
Don Cheadle	0	48
Marisa Tomei	0	48
Michelle Pfeiffer	0	47
Robert Downey Jr	0	46
Scarlett Johansson	0	44
Angela Bassett	0	43
Natalie Portman	0	42
Peter Dinklage	0	37
Chris Evans	0	33
Michael Douglas	0	33
Zoe Saldana	0	31
Jeremy Renner	0	30
Robert Redfort	0	29
James Spader	0	24
Michael Keaton	0	23
Betty White	0	20
Tom Cruise	0	20
Michael J Fox	0	11
Jennifer LoveHewitt	0	10

### Community 2

Id	Modularity Class	Degree
Christopher Walken	2	50
Charlize Theron	2	45
Matthew McConau...	2	41
Jennifer Garner	2	36
Kathy Bates	2	36
Anne Hathaway	2	34
Amanda Seyfried	2	29
Glenn Close	2	29
Leonardo DiCaprio	2	29
Sandra Bullock	2	28
Al Pacino	2	27
Keanu Reeves	2	27
Kate Hudson	2	26
Johnny Depp	2	24
Christina Ricci	2	22
Dakota Fanning	2	22
Sharon Stone	2	22
Diane Keaton	2	20
Julia Louis Dreyfus	2	13
Neil Patrick Harris	2	8

### Community 1

Id	Modularity Class	Degree
Meryl Streep	1	48
Ed Harris	1	45
Nicole Kidman	1	38
Winona Ryder	1	28
Reese Witherspoon	1	27
Jennifer Connelly	1	24
Claire Danes	1	21
Jared Leto	1	20
Renee Zellweger	1	20
KerryWashington	1	18
Gary Sinise	1	14

### Community 3

Id	Modularity Class	Degree
Julianne Moore	3	44
William H Macy	3	38
Sean Penn	3	36
Emma Stone	3	34
Naomi Watts	3	29
Joaquin Phoenix	3	26
Jeff Bridges	3	24
Chloe Grace Moretz	3	22
Halle Berry	3	21
Mel Gibson	3	21
Harrison Ford	3	20
Uma Thurman	3	19
JameEarl Jones	3	16
Sally Field	3	15
Helen Hunt	3	14
Jodie Foster	3	13
Keri Russell	3	12

## Community 4

Id	Modularity Class	Degree
Ben Affleck	4	48
Robin Wright	4	48
Bruce Willis	4	46
Matt Damon	4	46
John Malkovich	4	45
Morgan Freeman	4	43
Julia Roberts	4	38
Andy Garcia	4	37
Tommy Lee Jones	4	37
Amy Adams	4	35
Brad Pitt	4	35
Bryan Cranston	4	33
Rosario Dawson	4	33
Tom Hanks	4	33
Robert Duwall	4	28
Vera Farmiga	4	28
Joseph Gordon Levitt	4	27
Diane Lane	4	26
Drew Barrymore	4	25
Liv Tyler	4	25
Chris Pine	4	24
Denzel Washington	4	24
Jake Gyllenhaal	4	23
Jessica Chastain	4	23
Milla Jovovich	4	23
Kurt Russell	4	22
Viola Davis	4	22
Angelina Jolie	4	20
Will Smith	4	20
Viggo Mortensen	4	19

## Community 5

Id	Modularity Class	Degree
Susan Sarandon	5	60
Sanley Tucci	5	57
Woody Harrelson	5	49
Robert DeNiro	5	47
Steve Buscemi	5	44
Edward Norton	5	42
Willem Dafoe	5	42
Paul Giamatti	5	41
John Turturro	5	39
Bill Murray	5	38
Nicolas Cage	5	34
Frances McDormand	5	32
Danny DeVito	5	31
Jeff Goldblum	5	31
Harvey Keitel	5	29
John Cusack	5	29
Laura Linney	5	27
Salma Hayek	5	25
Adam Sandler	5	24
Jack Nicholson	5	22
Jennifer Lawrence	5	22
Alan Alda	5	21
Elisabeth Shue	5	18
Jessica Lange	5	17
Faye Dunaway	5	9

## Community 6

Id	Modularity Class	Degree
Pau Rudd	6	51
Elizabeth Banks	6	47
Mark Wahlberg	6	40
Kristen Wiig	6	39
James Franco	6	36
Steve Carell	6	36
Kirsten Dunst	6	34
Kristen Bell	6	29
Olivia Wilde	6	28
Jennifer Aniston	6	26
Mila Kunis	6	26
Melissa McCarthy	6	24
Sigourney Weaver	6	24
Tina Fey	6	24
Zooey Deschanel	6	24
Amy Poehler	6	21
Dwayne Johnson	6	20
Christina Applegate	6	18
Anna Kendrick	6	17



## 4. Structural Analysis

Structural analysis is the process of calculating and determining the effects of loads and internal forces on a structure. The results of the analysis are used to verify a structure's fitness for use

<b>Nodes</b>	150
<b>Edges</b>	2249
<b>Clustering Coefficient</b>	0,292
<b>Avg. Path Length</b>	1.815
<b>Density</b>	0,201
<b>Avg. Weighted Degree</b>	29,987
<b>Diameter</b>	3

In the network we have a total of 150 nodes in our network and As we mentioned before, the relationship between the actors/actresses was determined by whether there was a movie they starred in together or not, so as a result, 2249 edges were formed between nodes.

### 4.1 Average Degree

The average degree of the nodes in the network gives an idea of how dense a neighborhood is. If  $k_i$  is the degree of  $i$  th node, so for a network of  $N$  nodes total degree will be equal to  $\sum k_i$  where  $i = \{1..N\}$ . As for an undirected link between nodes  $u$  and  $v$  the degree is counted twice, so the total degree equals  $2 * \text{Links}$ . Hence  $k_{avg} = 2L / N$  for an undirected network.

## Degree Report

---

*Degree of common movie network;*

$$(2 * 2249) / 150 = 29,98$$

### Results:

Average Degree: 29,987

## 4.2 Clustering Coefficient

The clustering coefficient for the whole of the network is the average of the clustering coefficients of all the nodes. It is a measure of the degree to which nodes tend to cluster together. Also it can be defined as the measure of the friends of my friends who are also my friends. In this study this network shows low clustering feature according to the results,

## Clustering Coefficient Metric Report

---

### Parameters:

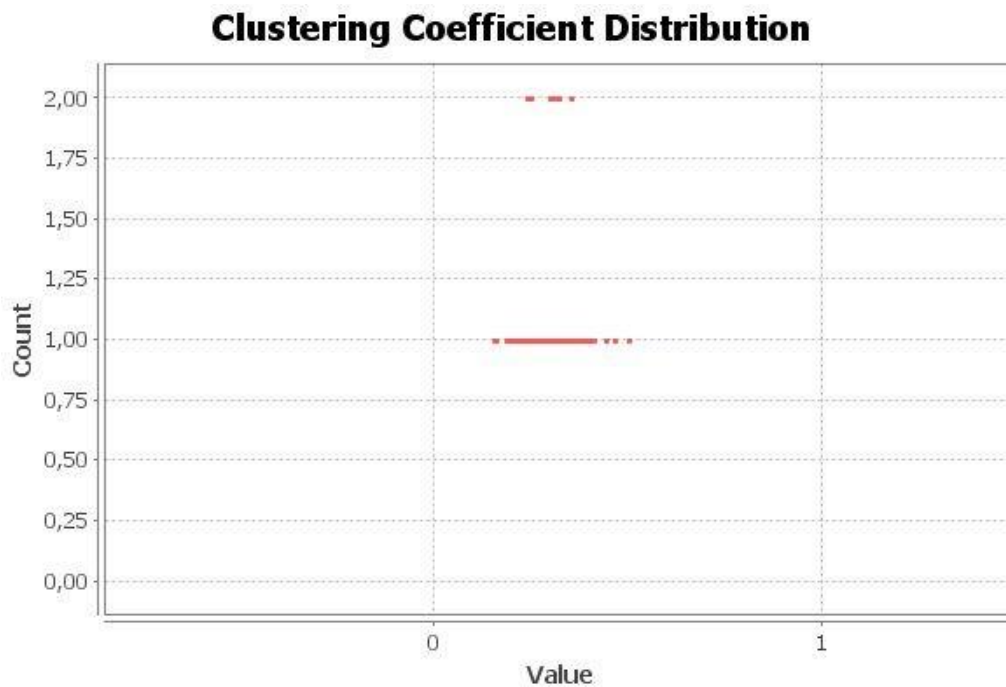
Network Interpretation: undirected

### Results:

Average Clustering Coefficient: 0,292

Total triangles: 7179

The Average Clustering Coefficient is the mean value of individual coefficients.



### 4.3 Average Path Length

Average path length is a measure of the efficiency or mass transport on a network. It is defined as the average number of steps along the shortest paths for all possible pairs of network nodes.

### 4.4 Density

Density is the ratio of the total number of edges in a network to the maximum number of edges that can be obtained by connecting all nodes with each other. Between two different networks with the same number of nodes, the one with a higher number of sides is denser. The maximum possible number of edges for a non-directional network is calculated as follows, where  $n$  is the maximum possible number of edges in a network with known node number:

$$\frac{n \times (n - 1)}{2}$$

In a non-directional network with  $n$  nodes and  $m$  sides, the density is calculated by the formula:

$$\frac{m}{n \times (n - 1) / 2}.$$

# Graph Density Report

---

## Parameters:

Network Interpretation: undirected

## Results:

Density: 0,201

For this graph;

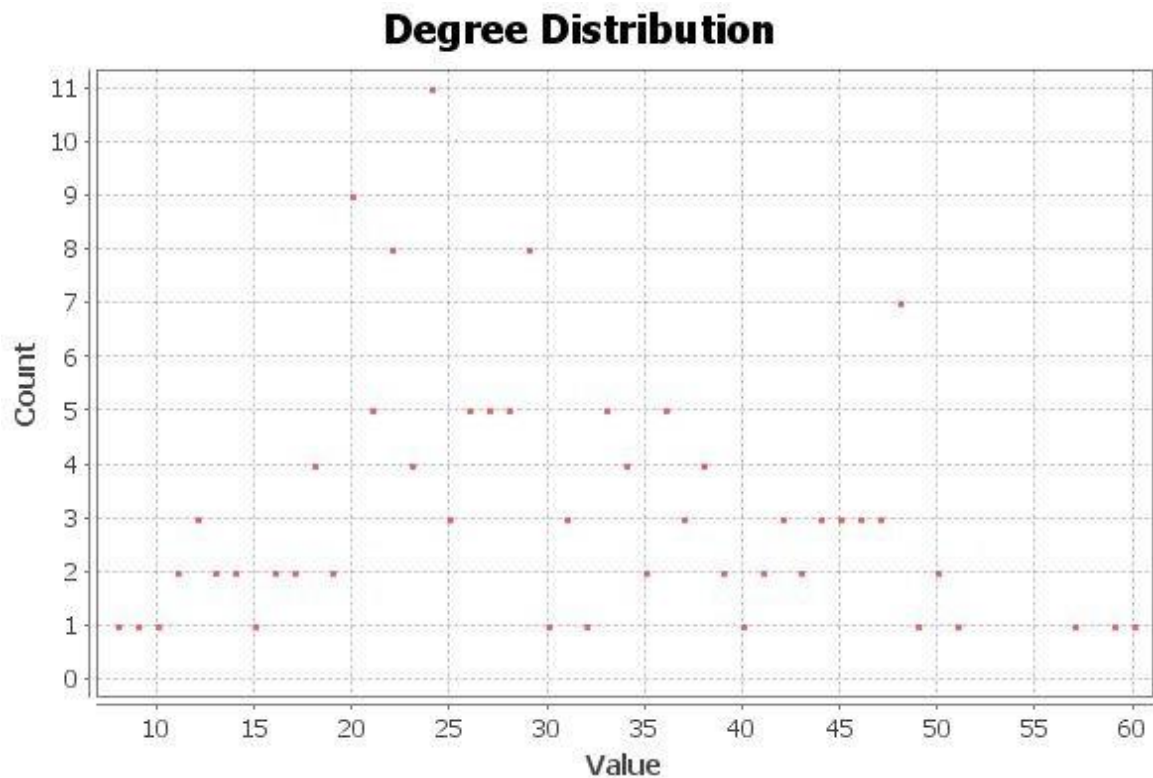
n = 150 nodes, m = 2249 edges

$$\frac{2249}{150 \times (150 - 1) / 2} = 0,201$$

## 4.5 Diameter

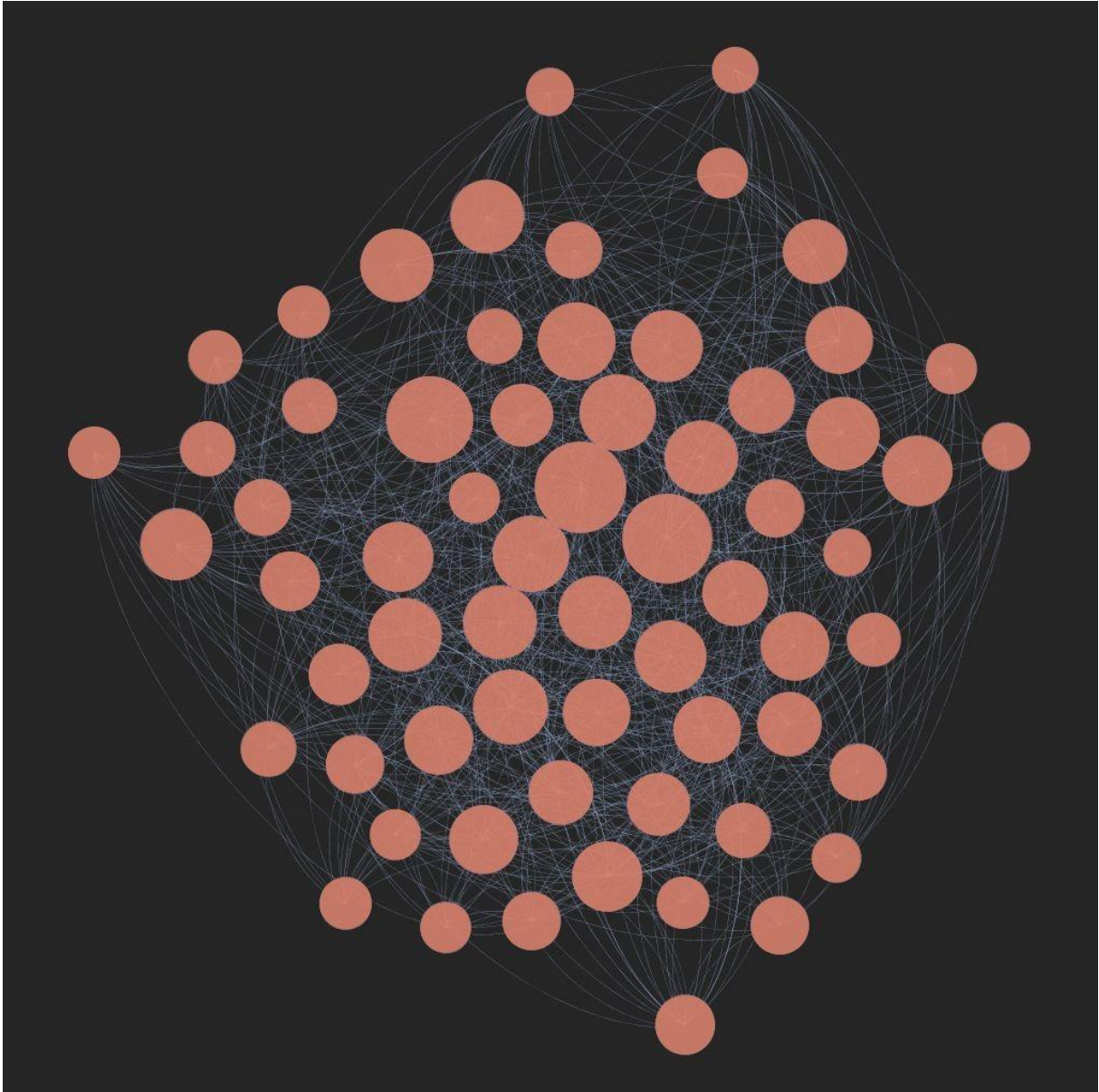
We can define the diameter of a network as the longest of all the calculated shortest paths in a network. Once the shortest path length from every node to all other nodes is calculated, the diameter is the longest of all the calculated path lengths. With nodes id of 1-2-3-4 are connected, going from 1->4 this would be the diameter of 3.

## 4.6 Degree Distribution



Degree distribution is a form of representation that graphically presents the distribution of degrees of changes in the network. The degree distribution is perhaps the criterion that gives us the most insight into the global structure of the network, as it allows us to understand how many nodes to which degree there are.

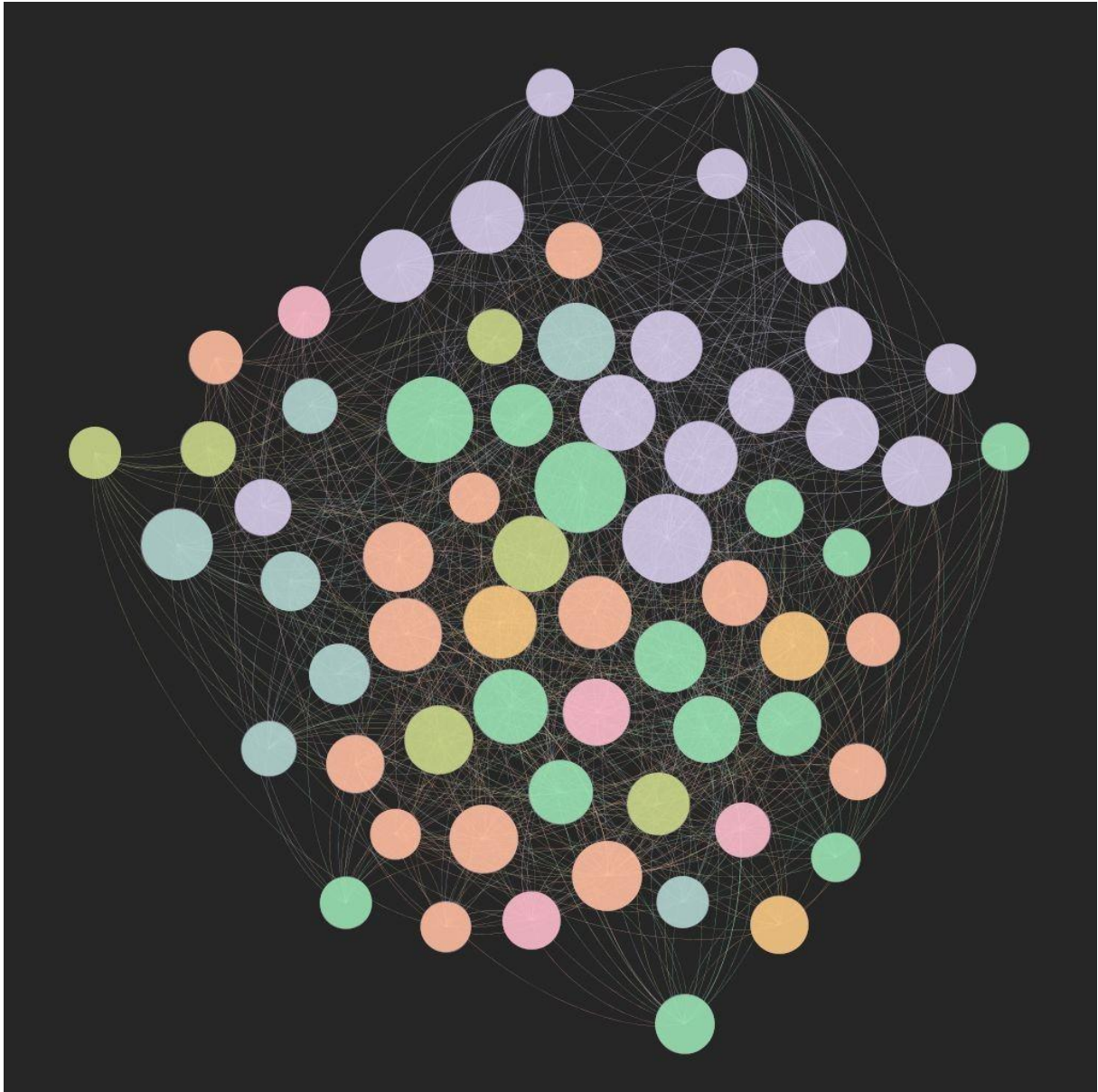
## 5. Filtering Samples of Our Network



**Figure 10:** *Nodes with degree greater than 30*

We filtered our network based on the nodes with its degrees greater than 30. The color of the nodes change according to the number of degrees the nodes have. As you go from small node to big node as well, the number of degrees the nodes has increases.





*Figure 11: Nodes with degree greater than 30 based on modularity*

We filtered our network according to the nodes with its degrees greater than 30 based on modularity. In the visualization of the network, the sizing of the nodes between 30-60 in proportion to their degrees and also the communities in the network are highlighted with different colors.