Best Picture
A Social Network Analysis

DNSC 6290: Group Project

Submitted By:
Edward Facundo
Elizabeth Freudenberger
Lawrence Gadsden
Ramandeep Kaur
Lingzi Kong

Introduction

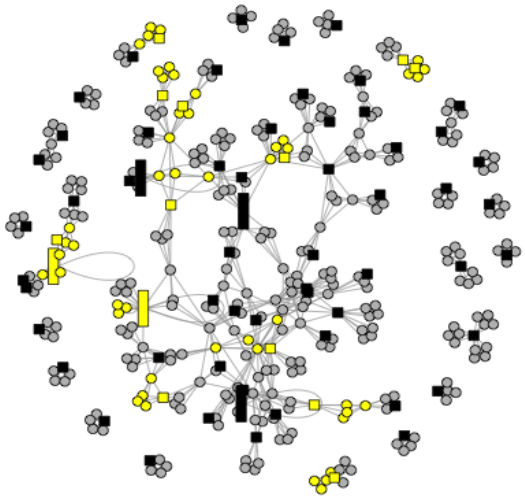
The Academy Awards, or "Oscars", is one of the most recognized annual movie awards in the film industry. [1] Oscar award winners, whether actors or directors, can gain fame and fortune by participating in a film that wins best picture. As the award season approaches, we wanted to explore the importance of connections among Hollywood actors and directors to create an Oscar-winning movie. In other words, how important are connections in creating a winning movie? Is there a winning recipe? Is the Oscar-winning community particularly close-knit? Through the use of network analysis, we sought to quantify the answers to these questions and shed light on the best picture network. Social network analysis is an effective tool that helps us understand connections between actors and directors in both the Oscar winning and losing community.

In order to capture the most robust dataset we sourced our network from the IMDB. The IMDB includes Oscar nominated movies from year 2000 to 2014 and provides information on each movie title including Oscar awards, actors (top four actors leading), director, genre, and release date. [2] Our analysis synthesized this information into a Hollywood network with both actors and directors as nodes and films as the edges. Through various social network analysis measures such as centrality, communities, and clusters attempted to answer those questions posed above.

Oscar Nominated Movies (2000-2014) - Network Level Analysis

Figure 1 depicts the complete network where circles and squares represent actors and directors respectively and black or yellow represents nominees and winners respectively. The black or yellow rectangles represent actor/director nodes which in some cases occurs when the actor looks for an additional challenge in life and decides to also direct. Connections, or edges, between nodes represent movies in which those actors and director participate. This figure also displays a major component in the middle along with other smaller-sized components that scatter its outer edges.

Figure 1.1: Networks among Actors/Directors in Oscar Nominated Movies from 2000 to 2014



[1] http://oscar.go.com/

[2] http://www.imdb.com/

After some evaluation we can determine our network contains 395 nodes and 1025 edges. This gives the network density a value of 0.01317, which indicates the probability of direct connections between any two nodes among Oscar nominated movies from 2000 to 2014 is relatively low. Due to the sheer size of our network, we would say this density is much lower than expected, as we believed the actor/director circle would be more dense.

Components

After looking at the network in its entirety we felt the need to break it down into smaller components to see if we could infer any valuable information from the structure of its parts. By using component analysis, we determined the network is made up of 24 components, ranging from a size of 5 to 231 nodes. Once broken into its parts we compared the percentage of winners in each type of component size to answer the simple question, does being a part of a larger or smaller component matter for winning?

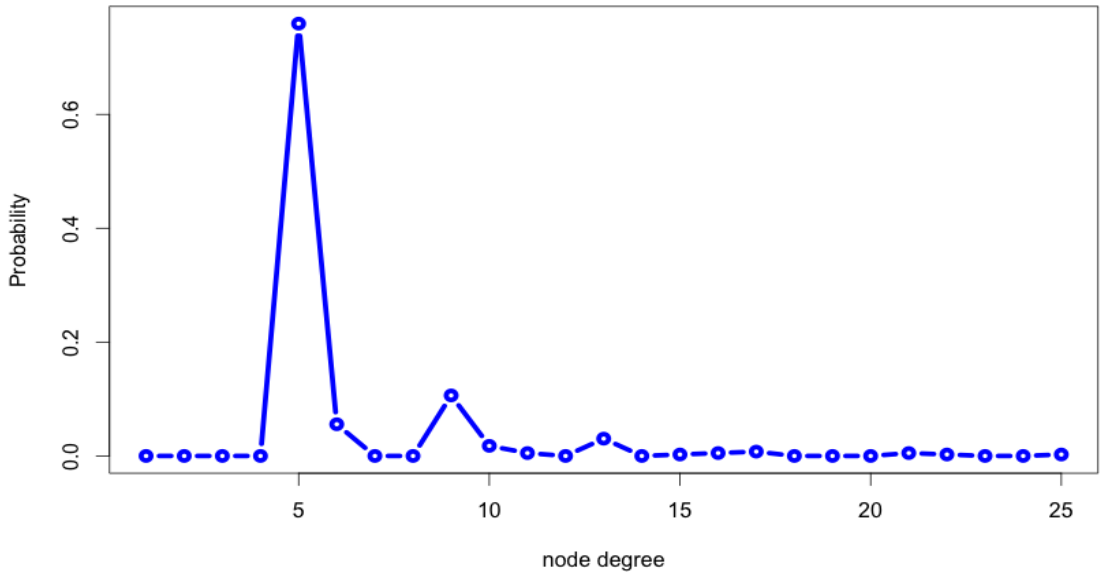
A quick evaluation of figure 2.2 tells us the vast majority of component in our network have a component size of just 5. This is likely due to the fact each movie in our dataset contains the top four actors and a single director making for a minimum of five nodes per film. Additionally, we can see, none of these small components have ever won the Oscar for best picture. While it cannot be exactly determined quite yet whether being in a large component helps a nodes' changes of winning, we certainly would want to be in a component with at least a size of 9 or greater.

Another interesting feature of this dataset is the relatively low win percentage of the largest node. With such a lower winning percentage, we may be able to infer that there is no evidence of a "winner circle" in which the same group of actors are winning.

Figure 2.1 Component Size Vs percentage of wins

Number of Components	Component Size (nodes)	% winners	*notes*
15	5	0%	
5	9	22%	
1	10	60%	*This is component #12. It is in the "northeast" part of the full network plots.*
1	13	0%	
1	21	38%	*This is component #8. See below. It is the grouping in the "west" part of the full network plots*
1	231	16%	*This is component #1. It is the main component in the center.*

Figure 2.2: Degree distribution of nodes



Network Centrality

In analyzing the various network centrality measures we broke our analysis down into two parts, Actors and Directors inside the largest component, and Oscar Winners vs. Losers. In order to address our question as to the winning formula, we needed to understand whether the various types of centrality had any sizable correlation with winners and losers.

Figure 3.1 below stratifies winners and losers by these centrality measures. Immediately apparent is the positive correlation in both degree and betweenness centralities between winners and losers. Specifically, degree centrality in the overall network, indicates winning nodes have on average 1.9 more Degrees than their counterparts. When comparing betweenness centrality, the difference is even more pronounced with winners scoring 300+% higher. Continuing our evaluation of figure 3.1 we can also see a variance in the outputs of the closeness and eigenvector centralities. We see that while the closeness measure seems to indicate little difference, the Eigenvector metric has a more pronounced difference of .015. As we know, the closeness measure of an overall network can be flawed due to its measure of the sum of geodesic distances rather than the eigenvector approach which uses factor analysis to identify the distances among actors within the network for its measure. Due to the size of our network, we believe the Eigenvector metric is a more telling measure of the closeness centrality within the network.

In addition to the three measures of centrality, we also investigated the average clustering coefficient. Like any high school freshman, we needed to determine whether or not cliques made a difference as to the popularity, or in this case probability of winning for a node. Interestingly enough, we discovered the opposite correlation between a node's likelihood to be a part of a triad and whether they actually won or not. An example we were able to see of this, was that out of the bottom 10 nodes ranked by the Clustering Coefficient, 6 of them were winners. This appears to be fairly good evidence that while important in some networks, the Clustering Coefficient in our network was less important in predicting winning nodes.

Figure 3.1: Centrality Statistics for Winners and Losers within full network

Was in a Best Picture Winner	Average Degree	Average Closeness	Average Betweenness	Average Eigenvector	Average Cluster Coefficient
Yes	6.790	0.0000119	679.421	0.054	0.700
No	4.892	0.0000115	197.303	0.039	0.903
Network Average	**5.190**	**0.0000116**	**272.977**	**0.041**	**0.871**

Figure 3.2: Centrality Statistics for Winners and Losers within main component

Was in a Best Picture Winner	Average Degree	Average Closeness	Average Betweenness	Average Eigenvector	Average Cluster Coefficient
Yes	7.211	0.0000151668	1100.268	0.087	0.683
No	5.368	0.0000151653	339.590	0.066	0.851
Network Average	**5.671**	**0.0000151655**	**464.723**	**0.070**	**0.823**

Is there a Winning Recipe?

After evaluating the general characteristic of our network, we wanted to see whether these metrics identified any particular nodes (actor/directors) with any consistency. Figure 4.1 offers us a view of the sorted and ranked nodes within our network ranked by each of our centrality and clusting measures. To our surprise, a single name came up as number 1 in each category, Leonardo DiCaprio. As it turns out, aside from being one of the most interesting actors in Hollywood today, he also happens to be the most centralized actor in the best picture network. With a Eigenvector Centrality of exactly 1, DiCaprio is positioned excellently within the largest component of our network. This means he has the best "reach" measure of anyone in Hollywood. In addition, he also has the highest number of degrees in our network, meaning he's been involved in over 24 films that have been nominated or won for best picture. Clearly his agent is doing something right. As our metrics show, it appears that if you want to be involved in a best picture, just follow Leonardo DiCaprio as he is in the right place, and has the right connections.

*Yellow = best picture winner

Table 4.1: Centrality statistics for important nodes

Rank	Degree Centrality	Betweenness Centrality	Eigenvector Centrality	Cluster Coefficient
1	Leonardo DiCaprio (24)	Leonardo DiCaprio (6175)	Leonardo DiCaprio (1)	Ben Affleck (0.107)
2	Brad Pitt (21)	George Clooney (5838)	Martin Scorsese (0.885)	Clint Eastwood (0.111)
3	Clint Eastwood (20)	Stephen Daldry (5319)	Brad Pitt (0.484)	Eli Roth (0.133)
4	Martin Scorsese (20)	Ed Harris (5032)	Cate Blanchett (0.444)	Bob Peterson (0.133)
5	Russell Crowe (16)	Sandra Bullock (4994)	Matt Damon (0.4)	Leonardo DiCaprio (0.134)
6	Sandra Bullock (16)	Ron Howard (4478)	Quentin Tarantino (0.398)	Brad Pitt (0.143)
7	George Clooney (16)	Kevin Bacon (4420)	Christoph Waltz (0.398)	Martin Scorsese (0.158)
8	Ethan Coen (15)	Sean Penn (4406)	Jonah Hill (0.355)	George Clooney (0.2)
9	Joel Coen (15)	Russell Crowe (4328)	Eli Roth (0.354)	Sandra Bullock (0.2)
10	Josh Brolin (14)	Jim Broadbent (4062)	Ethan Coen (0.332)	Russell Crowe (0.2)

How Close-Knit are the Winners?

Cliques

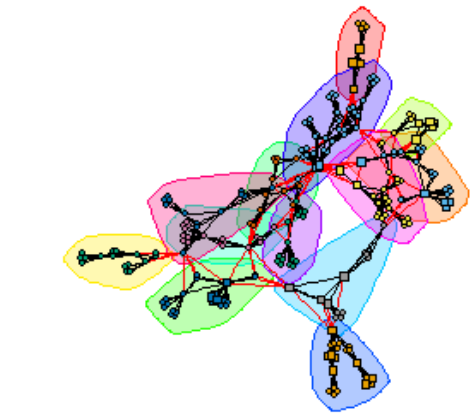
The largest component contains 1,930 cliques in total. The number of cliques is so high because all of the people that work on a movie are connected together! A normal movie containing four actors and one director has 31 cliques, while movies with two directors contain 63 cliques. These cliques are not significant and are simply an outcome of the data's structure. On the other hand, cliques that include actors and directors who are a part of different films would be important. However, there is no clear way to extract such cliques in igraph.

The largest cliques contain 6 nodes. There are 3 sets of these cliques within the largest component. Still, these cliques are not significant because they are all movies that have two directors. Hence, 4 actors and 2 directors produce a clique of 6 nodes.

Communities

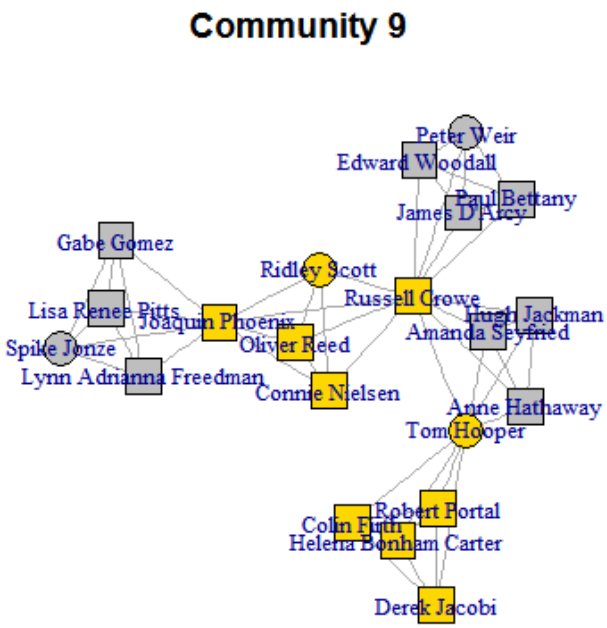
To find communities in the main component, we used igraph's edge.betweenness.community, a method that is based on the Girvan-Newman algorithm. In this method, edges with the highest betweenness are removed at each step.

Thirteen communities are found when using the Edge Betweenness method. The communities can be seen below in figure 4.1, while stats for each community can be seen in figure 4.2.



Community	Winners	Losers	Total	Percentage
1	5	4	9	56%
2	5	27	32	16%
3	0	16	16	0%
4	5	4	9	56%
5	5	17	22	23%
6	0	14	14	0%
7	0	17	17	0%
8	4	8	12	33%
9	**10**	**11**	**21**	**48%**
10	2	35	37	5%
11	0	15	15	0%
12	2	22	24	8%
13	0	13	13	0%

Out of the thirteen communities, Community 9 stands out the most because it contains the most winners (10) and these winners make up a significant portion of the community's nodes. When the community is looked at closely, one recognizes that the winning nodes are actors/directors from two movies that won the Best Picture award. Figure 3.3 shows Community 9. Actor Russell Crowe and director Tom Hooper are prominent members in the community.



Homophily

When testing for the homophily of winners (and losers) with the assortativity coefficient, the main component of the graph receives a result of .4983812. In many networks, an assortativity coefficient this high signals that homophily is high inside the network. However, for the Oscar's network, the number should not be taken alone because of the composition of the network. Again, every movie in the network is composed of four actors and one (or two) directors. All of the persons on the film work together. Thus, each winning node is linked to at least 4 other winning nodes. The connections within movies positively affect the assortativity coefficient so greatly that the network's homophily cannot be judged.

Conclusions:

- It is important to have connections in order to create an Oscar winning movie. Components with node size 5, meaning the nodes that did not have any connections outside of their own movie circle, did not win any Oscars. Thus, if you would like to win, you must participate in more than one Oscar winning movie.

- Our analysis shows evidence that winners tend to have higher degree and betweenness centrality. We observed that winning nodes have on average 1.9 more Degrees than their counterparts in both, the full network and the main component. Winners also had a degree of betweenness approximately 300% higher than non winners. Similarly, Eigenvector measures of winners were higher than losers by almost 40%. Therefore, a part of winning recipe is "Knowing people with good connections".

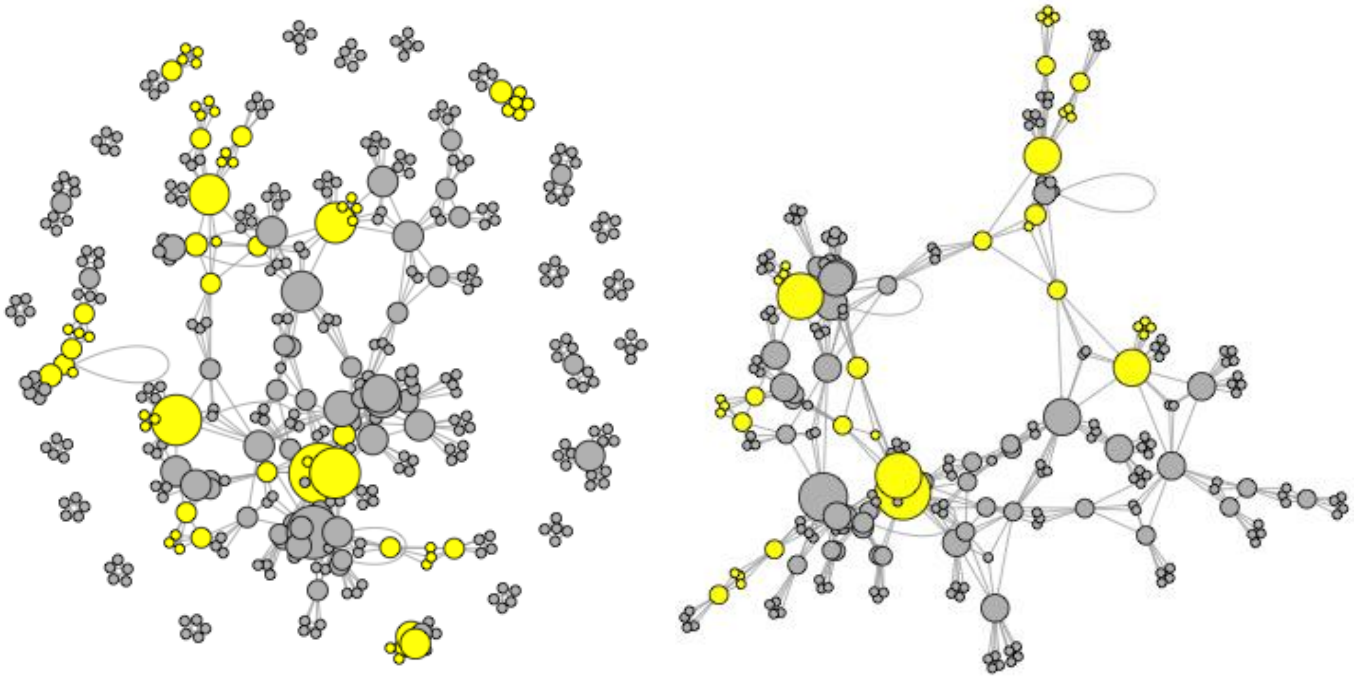
- Analysis suggests little correlation between the size of the component and the percentage of winning nodes. Mark Granovetter suggests the strength of weak ties may be more beneficial than direct connections attempting to influence a node. In this case, being a part of the largest component will help a nodes influence, however it may be the location within the component that actually matters.

- While the network's assortativity coefficient suggests winners tend to work together, the network's structure contaminates the test for homophily.

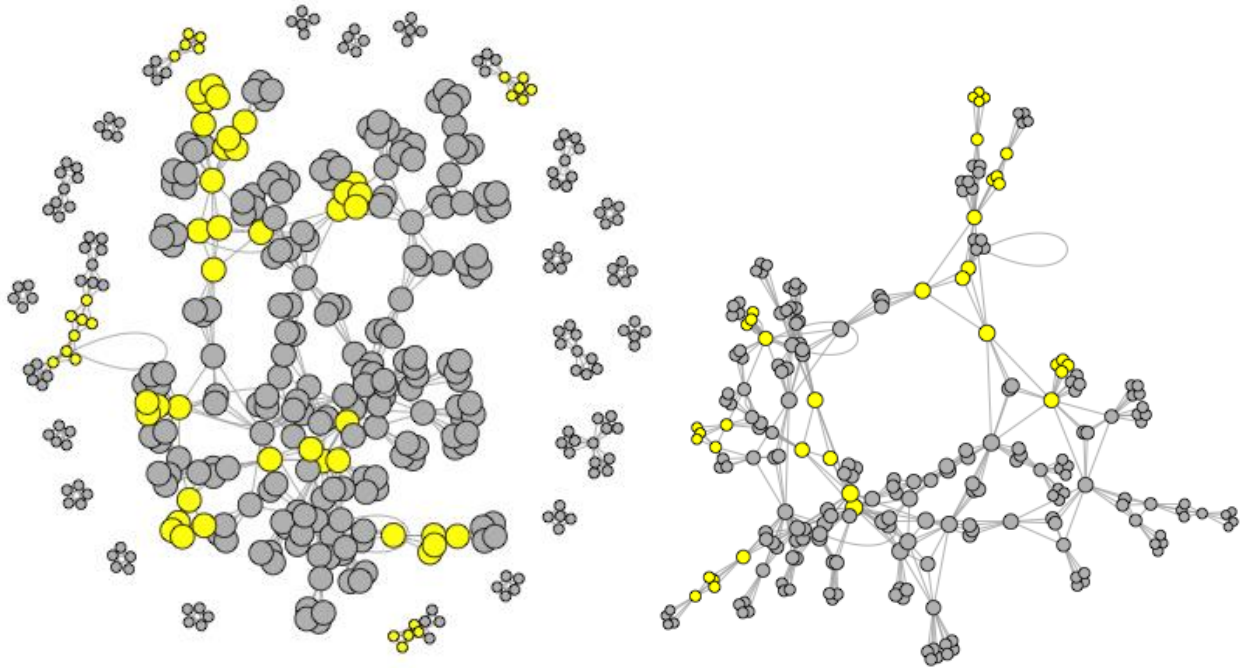
- Of all the actors and directors in our model, we predict Leonardo Dicaprio's would make a prime candidate for winning an Oscar as his centrality seems to consistently place him in the right movies.

Appendix

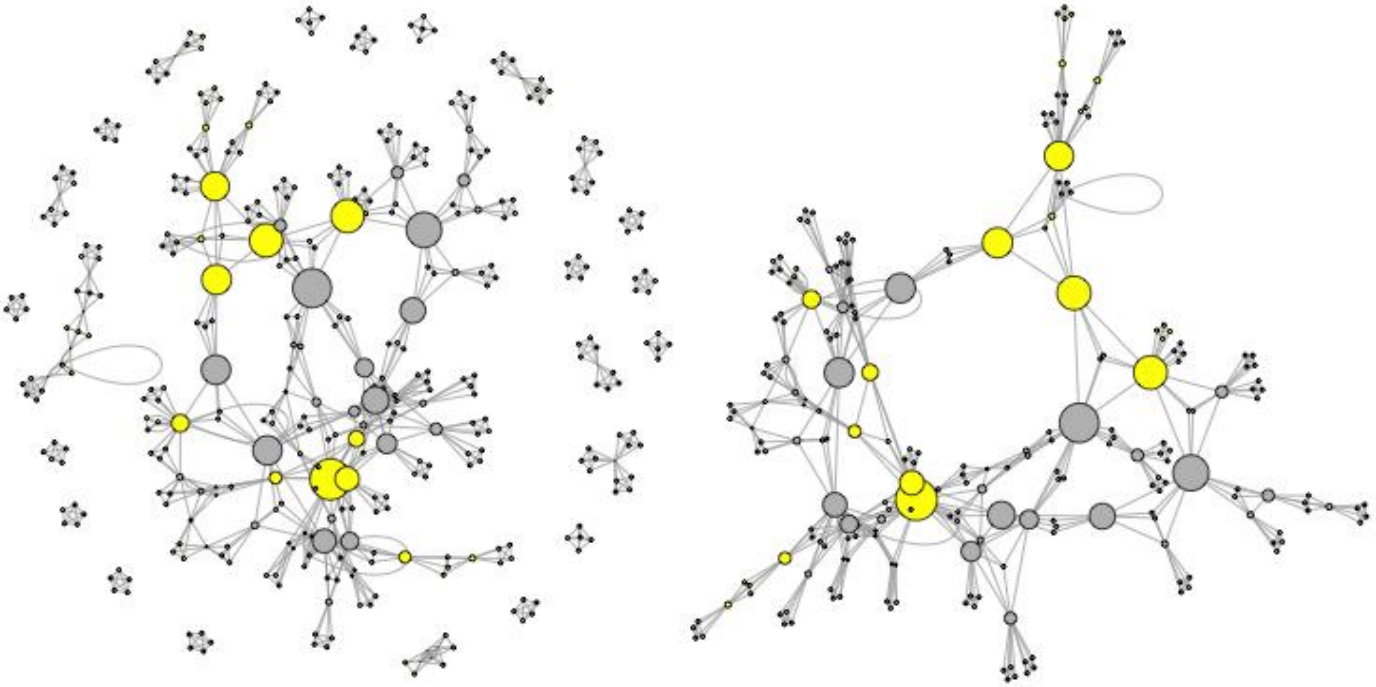
Degree Centrality



Closeness Centrality



Betweenness Centrality



Eigenvector Centrality