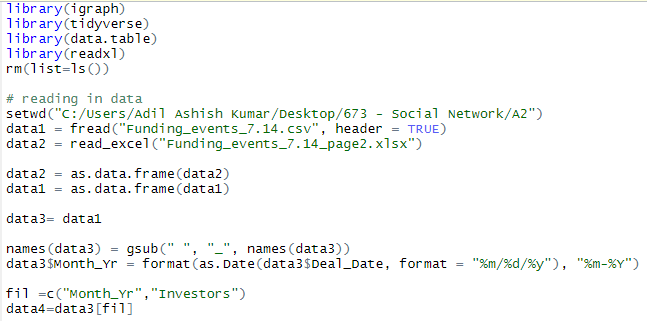
Name : Adil Ashish Kumar

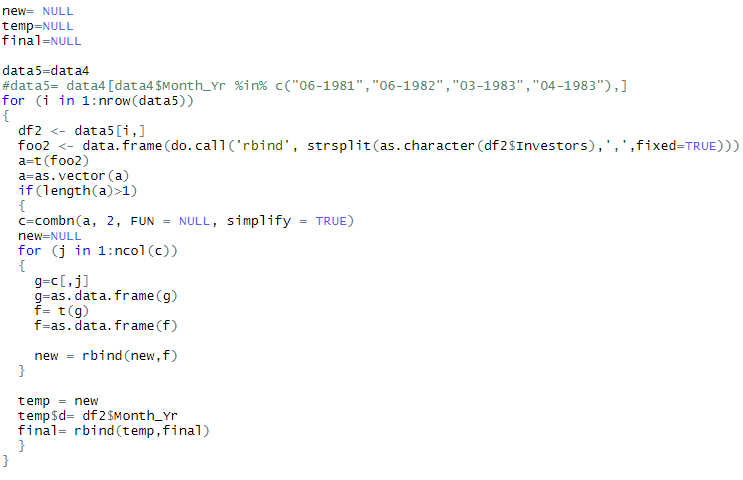
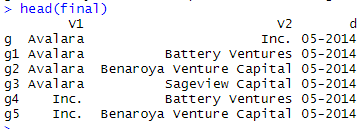
Emory MSBA 2019-2020

Social Networks ISOM 673 – Empirical Assignment 2

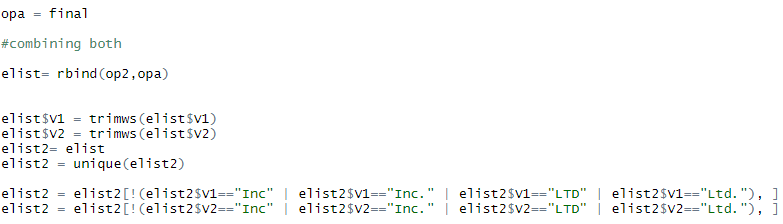
As a first step I did some data processing to convert the data into an edge list format. After reading the 2 funding event files, I first made sure the date formatting was consistent between the 2 files.



I then wrote a set of for loops to convert the data into an edge list. The logic was as follows: I first looped through each row of the data, split the investor column into multiple columns, and then transposed the same into 1 column. For this particular column I then found out all possible combinations of 2 investors. This was output as a set 2 investors, which I transposed and converted into a data frame with 2 columns along with the date. The final output was an edge list along with the date for each edge.

I ran the same loop for both event files. I know that this is not the most efficient approach, but I went with this since I could code it quicker. I first ran the loop on the smaller set to validate if my approach worked and then ran it on the bigger file. I also removed leading and trailing spaces from the investor names.



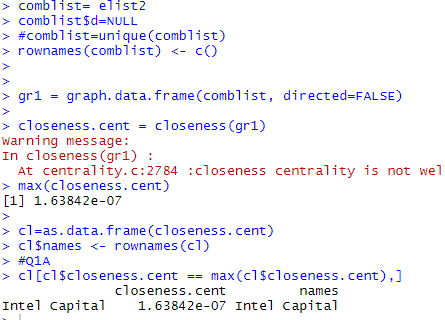
Looking at the edge list, I noticed that a few companies that did not really exist showed up – Inc and Ltd. This was because the investor column had commas between company names and words like “inc” and “ltd”. I decided to drop these rows since they did not make any sense.

1. First, perform the Kevin Bacon Hollywood Actor exercise on the venture capital firm network.

(A) Which firm is the center of the venture capital firm network as of July 2014? Consider

the most central firm to be the firm with the largest closeness centrality, as in the

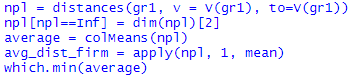
Hollywood Actor example.



For this, I took the final edge list and dropped the date column. I then created a graph out of this edge list and used the closeness function of the igraph package to calculate closeness centrality. Intel capital has the highest closeness centrality in the network. This means that it is the most central firm in the network. Closeness gives an idea of reachability in the network. Based on the output we can say that Intel Capital has high closeness and will tend to rely lesser on other nodes to reach different parts of the network.

(B) Next, compute the average shortest path length between all firms in the July 2014network and verify that the firm with the highest closeness centrality also has the lowestaverage path distance. You can consider nodes that are unreachable to be separated by a number of steps equal to the total number of the firms in the network.

For this I used the distances function from the igraph package to find the pairwise shortest paths between any 2 vertices. Based on the instructions I then considered those with infinite distances to have distance = no of firms. I took the average of these distances and noted that Intel Capital indeed had the lowest average path distance.



(C) What is the average shortest path length for all firms? Why is this number so high?

For this I used the mean\_distance function from igraph package. The value is 946.74. This value is high because it is probably affected by the higher path lengths of disconnected nodes. These unreachable nodes have been given path length no of firms, which is about 11000. So this is probably skewing the overall average path length.

2. Next, we will look at the development of the local group membership of the co-investment network over time. Allow the network to be updated monthly for each month t in the data, adding the new ties that occur through investments in the current month to be added to the existing network of ties that have occurred in previous months.

In Class Session 3, a figure on Slide 59 plotted over time the industry average of the highestdegree k-core each venture capital firm in the co-investment network belonged to. When a node is a member of a k-core with a high degree, its surrounding ties are very dense. When many nodes are members of k-cores with high degrees, this suggests that there may exist dense clusters within the network.

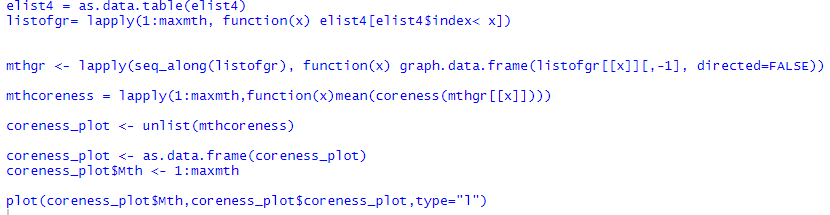
(A) Construct a figure similar to Class Session 3’s, plotting the average k-core of each venture capital firm in the network over time. This can be computed using the igraph function coreness. On the x-axis should be time. On the y-axis should be the highest-degree k-core each venture capital firm belongs to, averaged over all firms in the network up

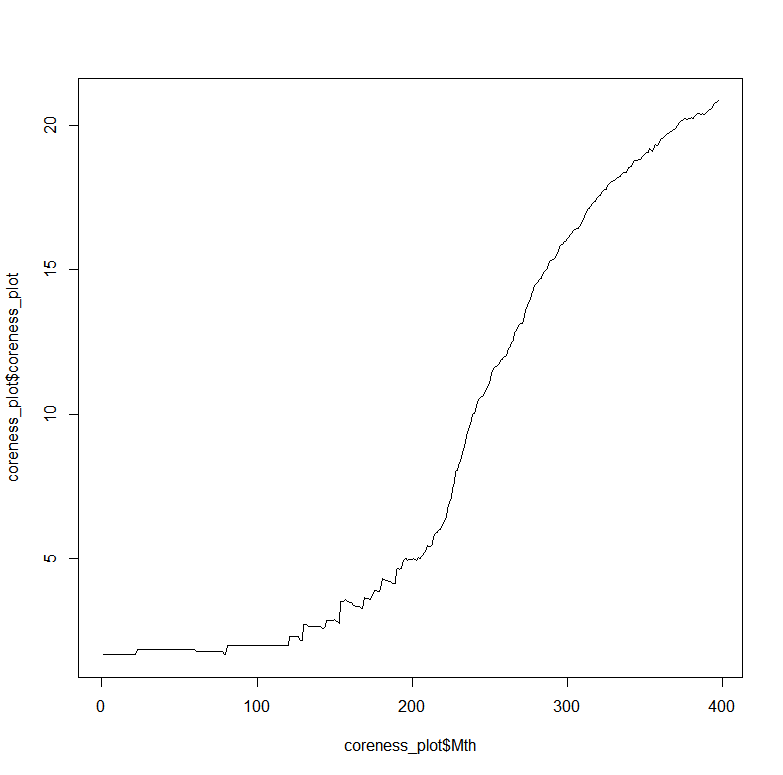
to that month.

For this question, I first generated a sequence of dates using the seq() function from lubridate. This helps number the dates in a sequence. I have given the sequence in ascending order of dates. I then joined this sequence back to my edge list so that I can create date wise graphs.



In this case, I considered the entire graph with multiple ties across years. I first used the lapply function to list of data tables for each month. Each data table for a month had the edge list for that month. This will be useful since we need to calculate monthwise networks. Next I used the lapply and graph.data.frame function to convert the earlier list of data tables into a list of graphs, one for each month. In order to calculate coreness, I used lapply to calculate the coreness for each graph. I then extracted the coreness scores and plotted them over time.

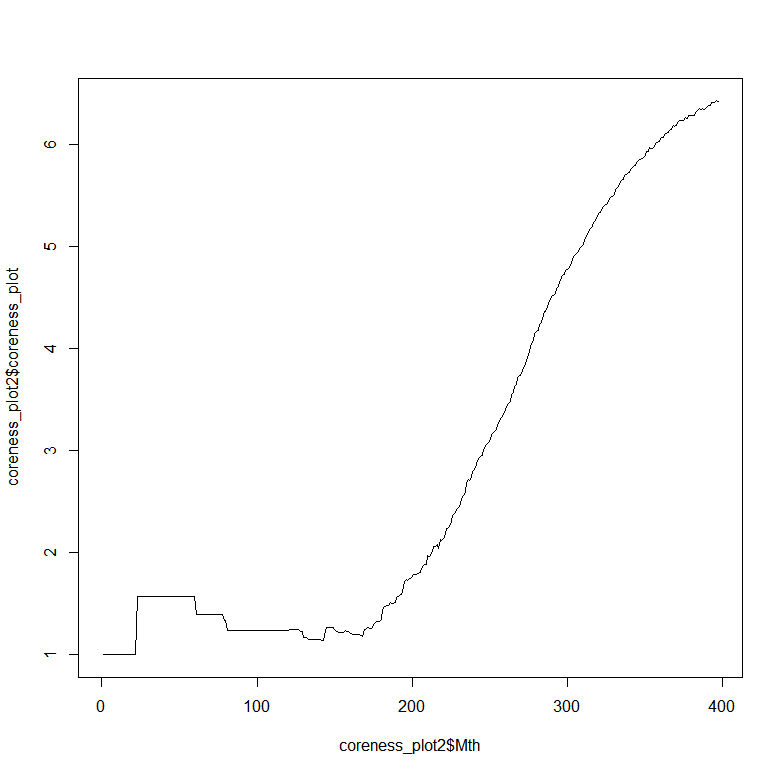




As we can see from the above plot, the coreness score does not change significantly initially for the first 100 days. Then we see a sudden increase in coreness as months go by. This increase is consistent with time. Higher the coreness indicates that more nodes have a higher degree. This makes sense, as the months go by, investors make more and more investments with other firms , growing their network at the same time by creating ties with newer firms

(B) Construct a plot similar to (A), but only consider unique ties as opposed to repeated ties in the calculation. Does the figure appear different than before? What does this suggest about the nature of relationships in the co-investment network?

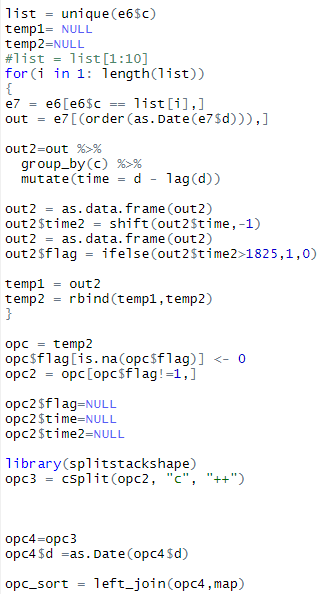
In this case, in order to remove repeated ties, I used the simplify function from igraph with conditions remove loops and remove multiples= TRUE. This will make sure that the graphs do not have repeated ties between edges. I then calculated the coreness for each graph using the lapply function on the graphs for each month. Below is the plot



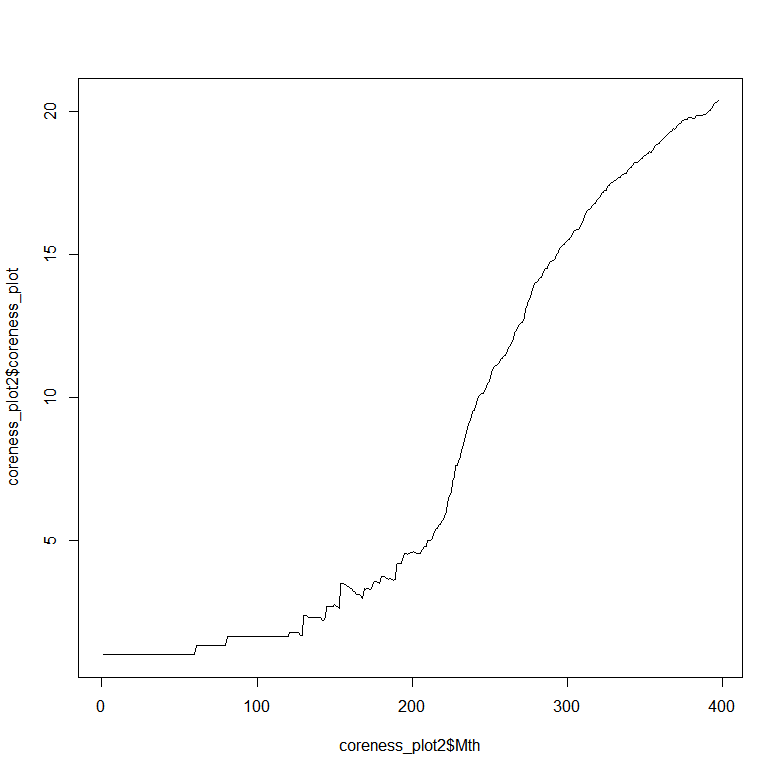
In this plot we can see a slight difference in the coreness values in the initial months. It is important to note the difference in scale of this plot and the previous one as well. We can see that overall coreness scores are on a smaller magnitude in this plot versus the previous plot. This makes sense since this network has fewer ties than the previous network, and hence has lower coreness scores. The lower coreness score indicate that the nodes have a lower degree , which again makes sense due to the fact that this network has lesser ties.

(C) Construct a plot similar to (A), but now allow ties to “decay.” Remove ties from the network if they are not renewed within 5 years. Does the figure appear different than before? What does this suggest about the nature of relationships in the co-investment network?

In order to filter out ties which were decaying, some additional data prep is needed. I used a for loop to loop over each tie. Within each tie , I used lead and lag variables in order to calculate the decay time for each tie. For any ties greater than 5 years that were not renewed I flagged them as 1. I then filtered out these ties and created a graph out of the same.



Once the data prep is done, I used this new edge list to create graphs. I followed the same process as previous 2 cases to plot the coreness scores by month. First I created a list of data tables for each month using lapply. I then created graphs for each months using the lapply function and graph.data.frame function. I then calculated the mean coreness for each graph by month. Below is the plot



Looking at the plot the overall trend of coreness is the same. Again it is important to note the magnitude of the coreness values. We can see that it is much higher in this case than the previous case. This makes sense since in this case repeated nodes are allowed as long as they are renewed every 5 years. The magnitude of coreness values is similar to the first case where all ties were considered. This indicates that most ties were renewed every 5 years, resulting in high coreness values, which indicate high degree of nodes in the graph.

3. Next, we will look at the development of the venture capital firm co-investment network in terms of its global core-periphery structure. Allow the network to be updated monthly, as in Question 3, but only consider the network that takes into account tie decay.

(A) Use the co-investment network’s concentration to determine if it tends towards a core periphery structure over time and demonstrate this visually. Begin the analysis after the very early period of the data when all of the firms have the same eigenvector centrality.

Illustrate a plot showing the maximum concentration score for each month of the data.

Unfortunately due to lack of time I could not execute this problem. However, this would be my approach to the solution.

First I would find out at what period in the data is the point where all firms have same eigen vector centrality. I would then attempt to calculate concentration scores for each month in the data. For this I would need to calculate the correlation between coreness scores and ideal coreness scores. Since the coreness scores are already calculated as part of the previous problem, I would use similar approach to calculate coreness scores and ideal coreness scores. The ideal coreness scores would be a binary variable based on the continuos coreness scores. I woud use a certain cutoff, maybe the mean coreness score to calculate the ideal coreness score. I would do this similarly for each month and calculate the concentration scores by month. I would then plot the maximum of concentrations scores for each month across all months.

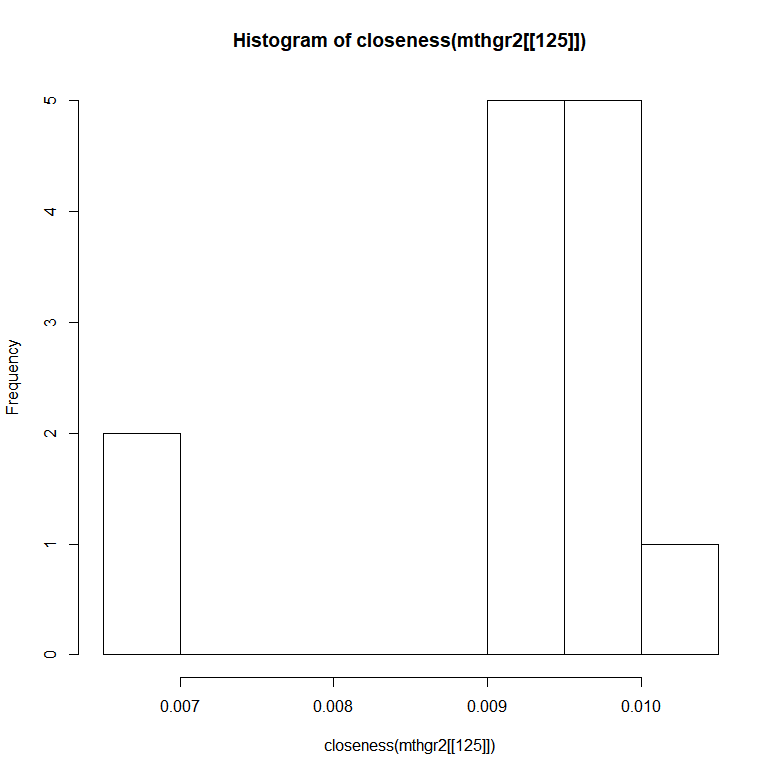
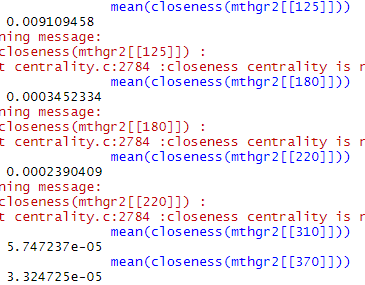
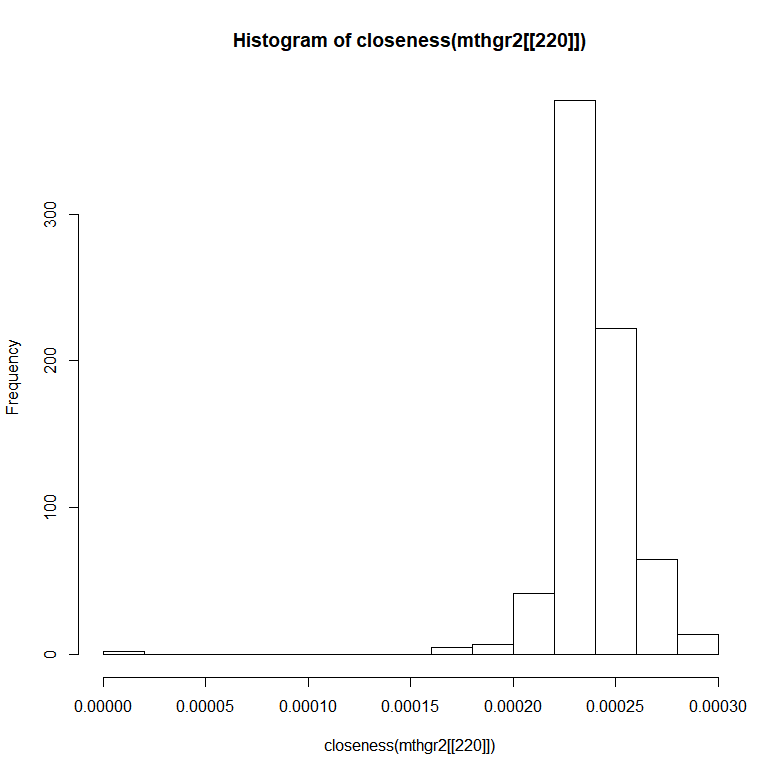
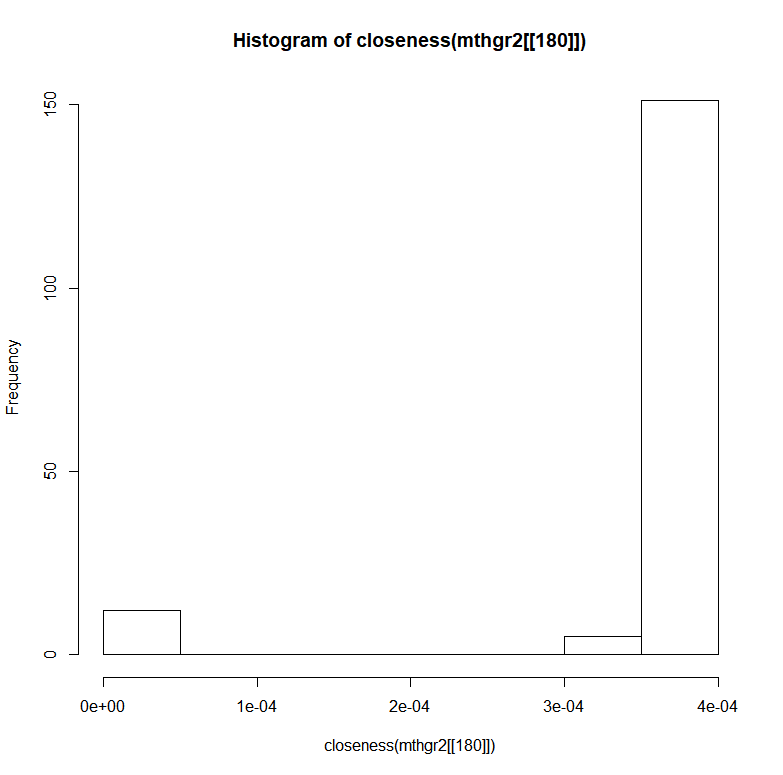
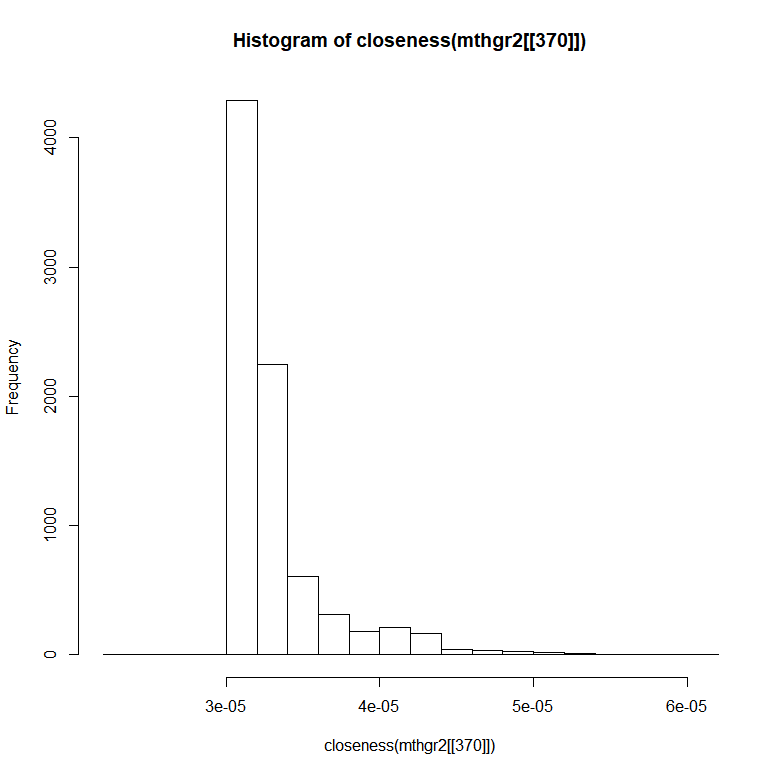
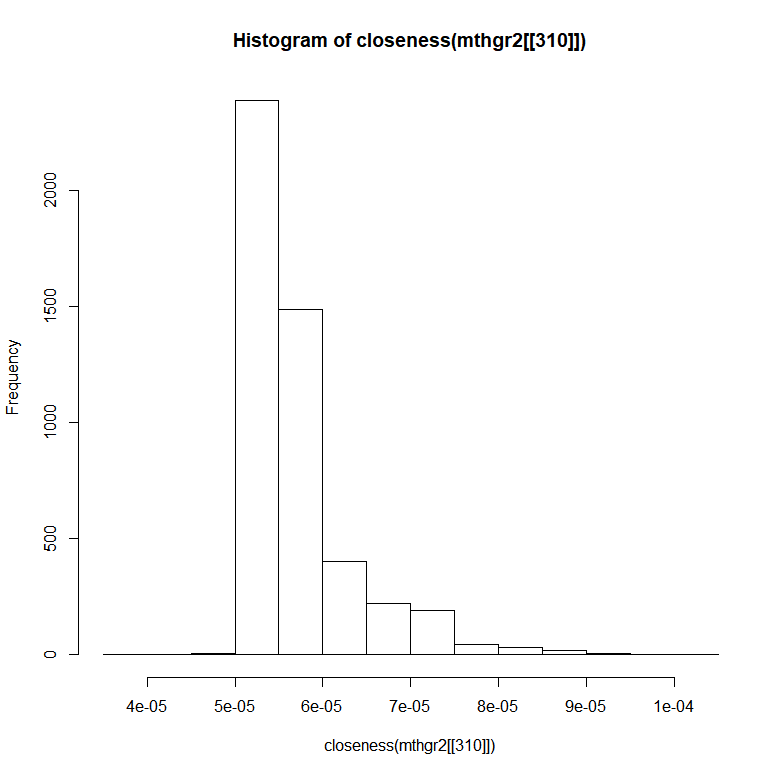
Illustrate a plot showing the proportion of firms in the ideal core partition corresponding to the maximum concentration score for each month.

In order to distinguish between firms in the ideal core partition and those outside, it is important to find the best fitting core with a moderate no of nodes. This would provide enough evidence that the network tends to a core periphery structure. I would observe to see if the best fitting curve contains all or only 1 node, in which case it maybe better to use a different structure to explain the network. Based on the ideal coreness scores, I would categorize firms into those that belong to the ideal core partition and those that do not belong. I would then plot the proportion of these firms belonging to the ideal core partition against max concentration scores for each month. The ideal core partition represents firms that would ideally be all connected to one another.

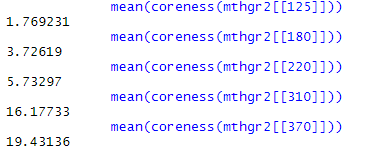
Illustrate a figure, with one plot for a month from each calendar year in the data,that shows the range of concentration scores for each partition size p in the network for that month’s snapshot.

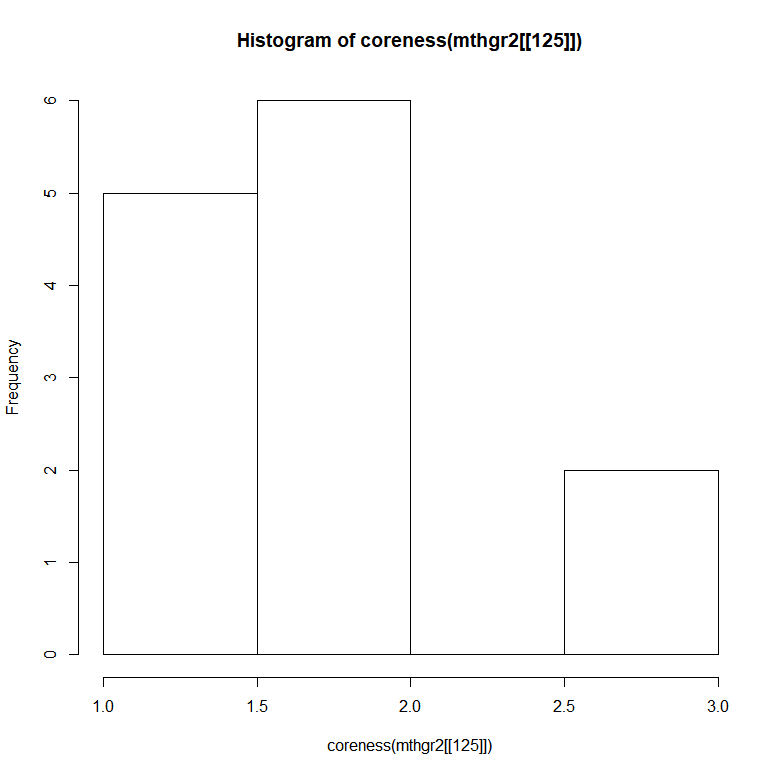
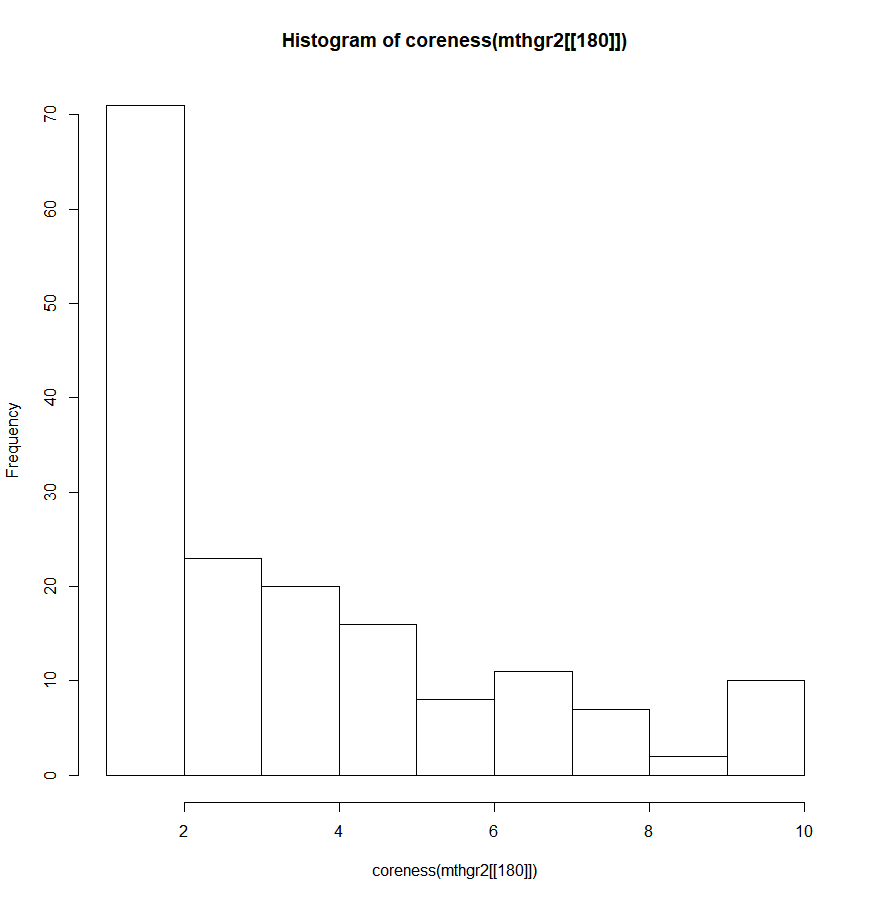
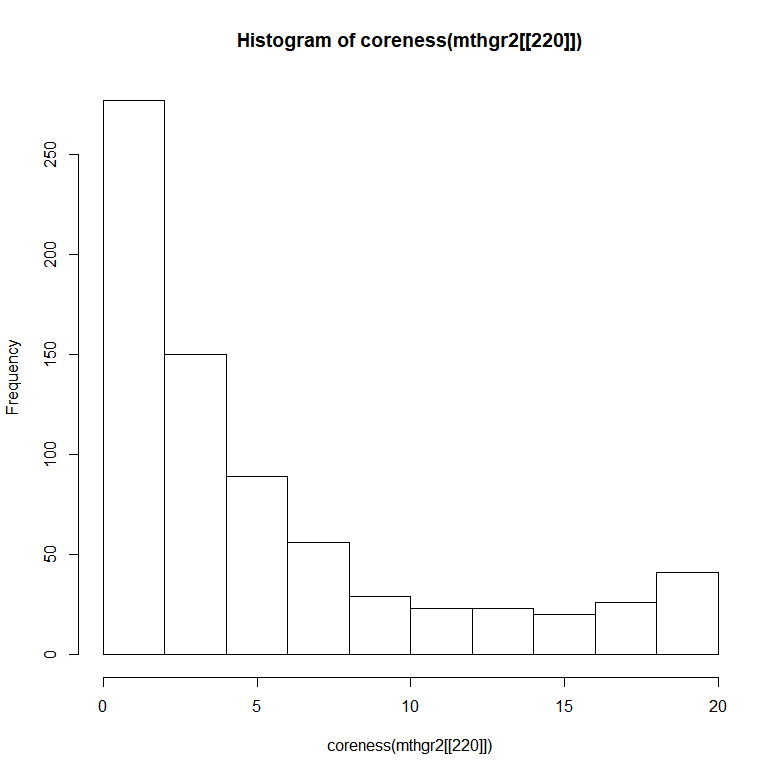
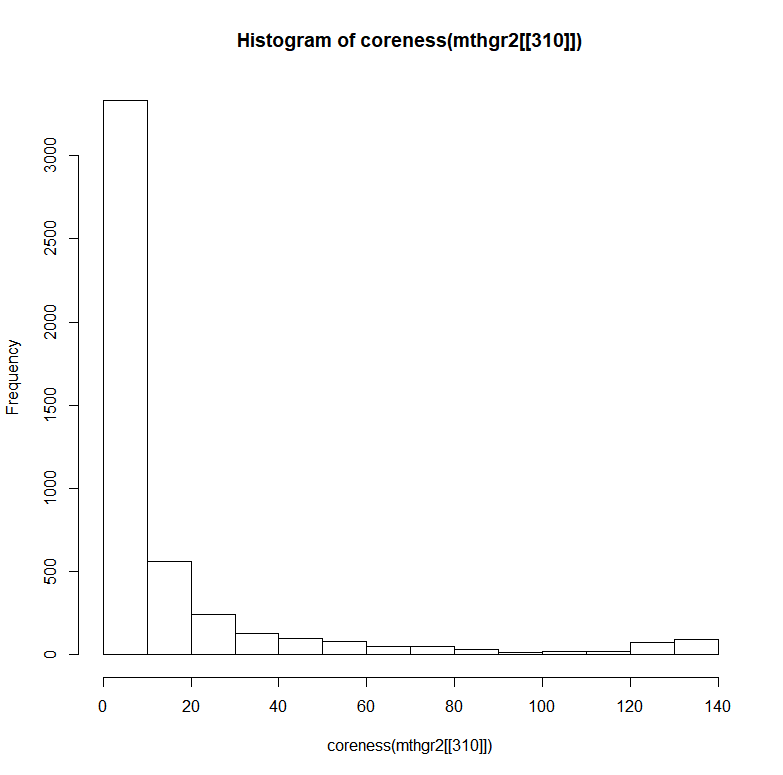
In order to find potential cores, we would divide the network in to core of size p with range 1 to n nodes in network. Within each partition, we could determine the range using max and min of concentration scores by p for each month in a year.

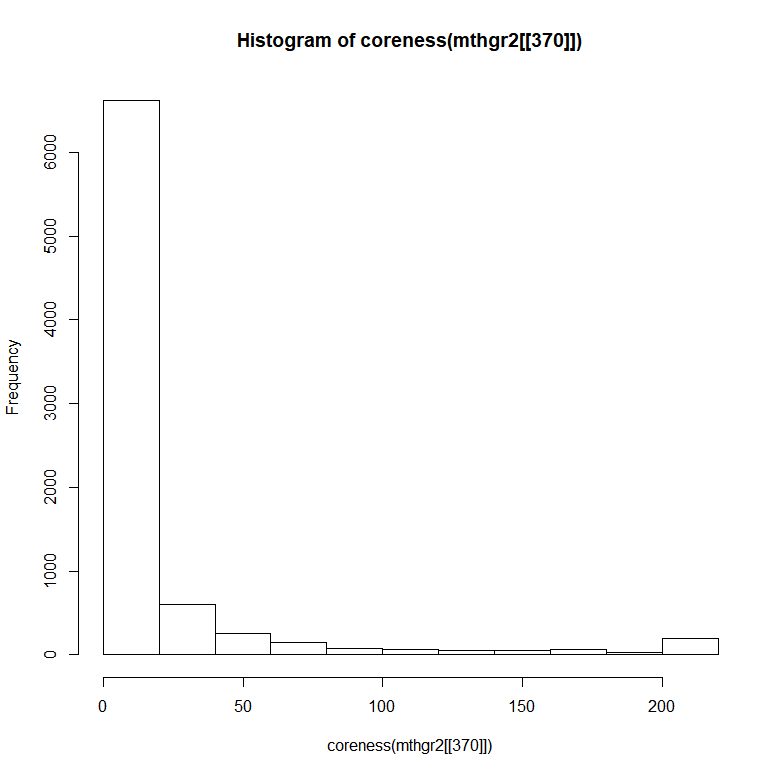
(B) Do you think that the recent network now exhibits more of a core-periphery structure or a structure made up of distinctly clustered components? Provide two other pieces of descriptive evidence outside of the conentration scores to support your conclusion.

The concepts of closeness and coreness also give an idea about network structure. Above I had a look at mean values and histogram of graphs for certain months over time. From the mean values we can see that closeness reduced as no of months passed by. From the histograms we can also see that the overall magnitude of closeness reduced as no of months passed by. We can also see from the histograms that the closenes values became slightly lesser skewed with time, indicating that closeness values became smaller and similar over time. The closeness centrality gives an idea of reachability in the graph and nodes with stronger closeness need not depend on ther nodes for reachability.





For the same months, I also looked at the coreness values. Looking at the mean coreness values, we can see that coreness is increasing with time. From the histograms we can see that the overall magnitude of coreness is increasing with time. Higher the coreness, higher the degree of nodes in the network.

Looking at both closeness and coreness it seems like the network structure is more like a core periphery structure rather than a cluster structure. This is because the closeness is reducing and coreness is increasing with time, forming a strong core rather than distinct clusters.

4. Last, we will analyze whether being in the core, being at the center of the network, and being a member of a densely connected group helps venture capital firms and the entrepreneurs they work with to perform better. You may use whichever statistical approach you wish to determine the direction and strength of the relationship between network position and a venture capital firm’s performance.

1. Is a venture capital firm being in the core, being at the center of the network, and being

a member of a densely connected group of the network related to having more more

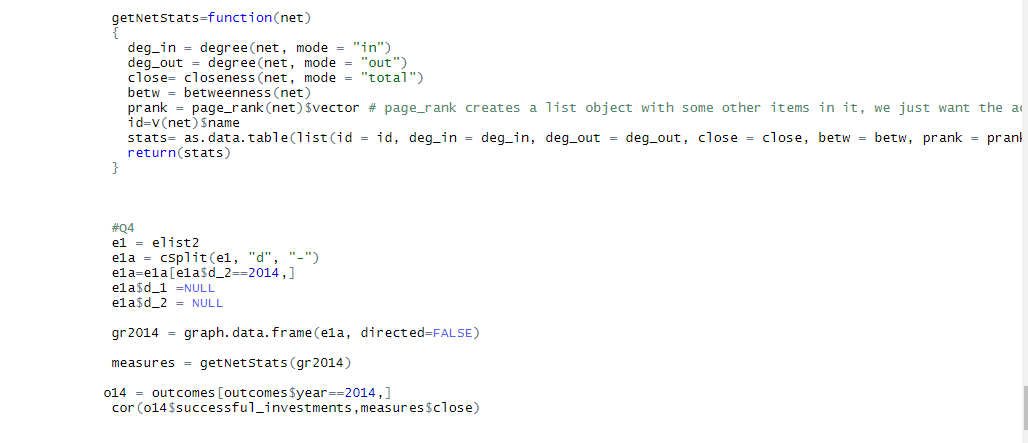
successful investments in a given year?

The outcome variable of successful investments is a non-negative integer, so the count

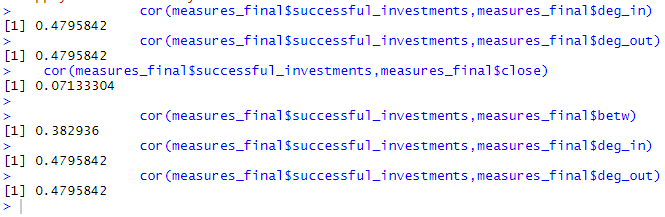
family modles can be useful. Some approaches are described at https://cran.r-project.

org/web/packages/pscl/vignettes/countreg.pdf.

For this I used the approach of finding centrality measures of a graph and checking correlations between outcomes and centrality measures.



I used a function from last homework to calculate centrality measures for a graph. I then used this function on a graph for the year 2014. I then ran correlations between the centrality measures and outcomes. I used correlations in order to see if a firms centrality statistics had any relation to its performance.



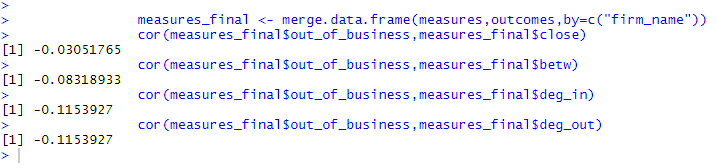
Based on the correlations we can see that higher the degree, better the chances of successful investments. Even though the correlation values are not very strong, it is still moderate and strong enough to indicate that a firm with better connections and betweenness has more successful investments. This also makes sense since a firm with more ties would generally be more aware of the market and have more opportunities for investments versus a firm with lesser connections. Degree centrality indicates measure of communication activity and frequency, which definitely ties up in this case for firms. Betweenness indicates measure of bridging and brokerage, which would allow firms to get in contact disconnected firms and realize better opportunities. Thus it does look like being in the center of a network or being in a connected network might increase chances of successive investments

(B) Is a venture capital firm being at the center of the network related to being less likely

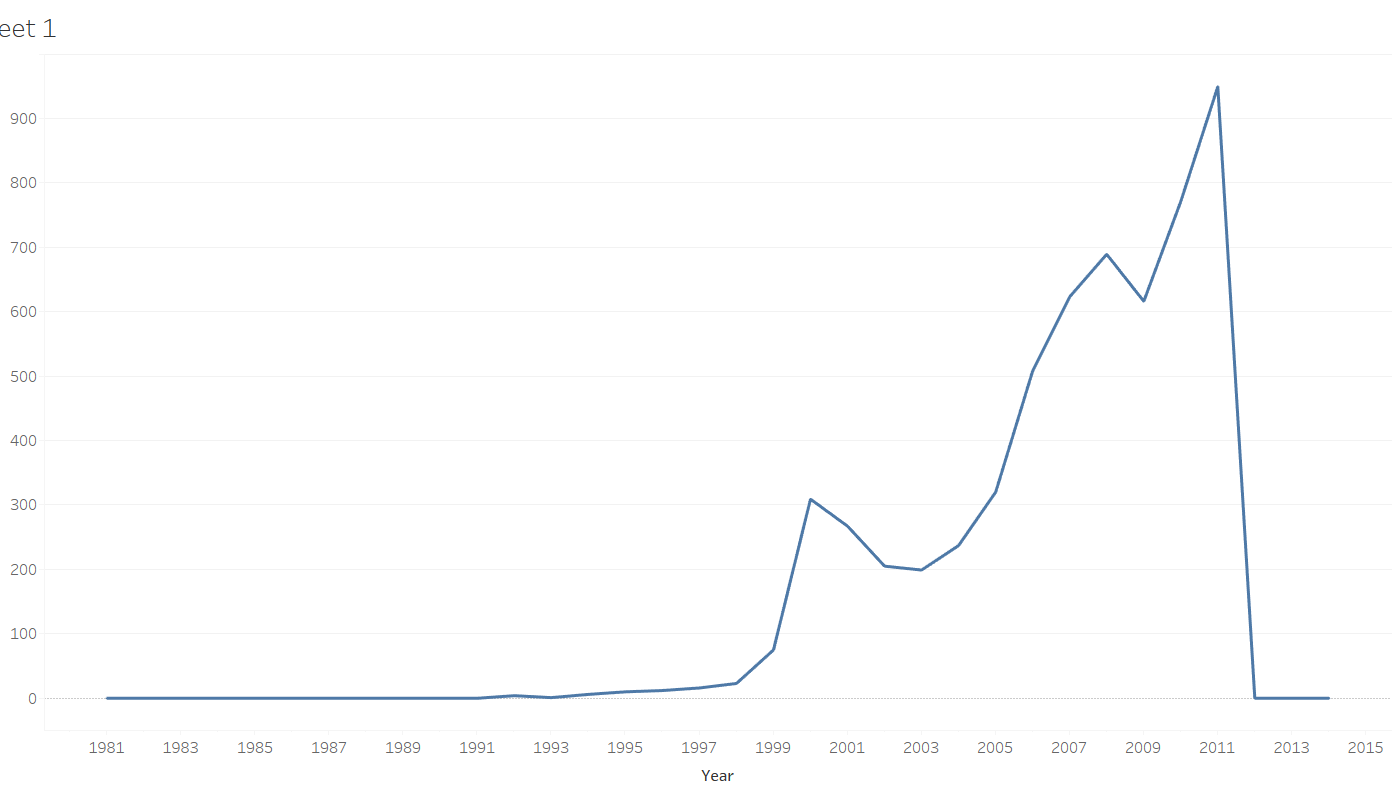
to go out of business?

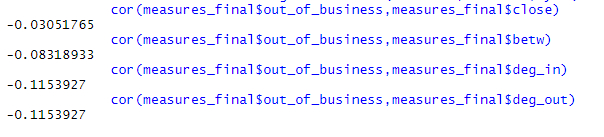
The outcome variable of going out of business is an event that can happen once, and the likelihood of this event depends on how long a firm has been in business. As a result, the survival family of models can be useful. Some approaches are described at

https://www.r-bloggers.com/survival-analysis-with-r/.



In this case, I decided to run correlations of the centrality measures against out of business variable. An interesting point to note is that the years 2012-2014 had no firms going out of business. That is certainly strange given the trend.





I decided to pick the year 2001 and see the correlations of centrality measures vs out of business. From the correlations we can see that all of them are negative and have a low magnitude. While the magnitude is low, the direction does indicate that out of business is inversely related with strength of centrality measures. This does make sense intuitively, as the lesser no of investors a firm is connected to , the lesser money it might be able to raise. Thus it does look like being in center of a network or in a strong connected network makes it less likely to go out of business.