Chad Germany, Xinhao Liu
Stat 480: Big Data Analytics
November 4, 2023

Homework 4

We'll explore casts for 'drama' movies from 1980-1999. I have limited the data to actors in more than ten productions over this time period (and to movies with more than ten actors).

Question 1. The actors network has an edge if the two actors were in the same movie. Plot the entire actors network.

FIGURE 1. displaying the actor network.

```
1
2 # Question1
3 plot(actnet, vertex.color="pink" , vertex.label=NA, vertex.size = 5)
```

LISTING 1. R source code implemented for the homework.

Question 2. Plot the neighborhoods for "Bacon, Kevin" at orders 1-3. How does the size of the network change with order?

```
1
2 #question 2
3
4
5 kevin_node <- "Bacon, Kevin"
6
7 neighbor_hood <- graph.neighborhood(actnet, 2, V(actnet)[kevin_node])[[1]]
8
9 V(neighbor_hood)[kevin_node]$color <- "gold"
10
11 V(neighbor_hood)$label.color = "black"
12
```

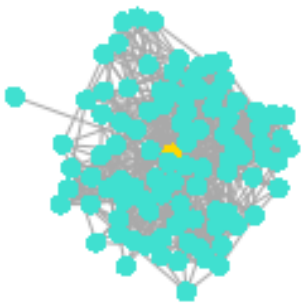

FIGURE 2. Network where the order is equal to 1 or nearest neighbors.

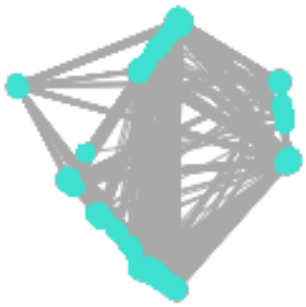

FIGURE 3. Network where the order is equal to 2 or next nearest neighbors.

```
13 V(neighbor_hood)$frame.color = NA
14
15 plot(neighbor_hood,  vertex.label=NA,edge.curved=FALSE)
```

LISTING 2. R source code implemented for the homework.

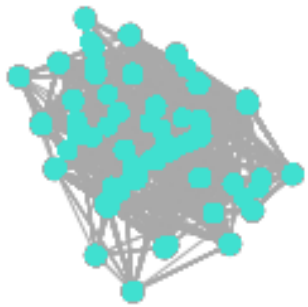

FIGURE 4. Network where the order is equal to 3 or next next nearest neighbors.

We see in the first order graph figure 2 there's a great concentration of nodes close to the center. This is what you would expect for nearest neighbors. For figure 3 we have nodes shown to be further from the center so less dense. These nodes show the next nearest neighbors. Lastly figure 4 shows next next nearest neighbors which are so far away that the center cannot be seen.

Question 3. Who were the most common actors? Who were most connected? Pick a pair of actors and describe the shortest path between them.

The most common 10 actor and actresses are Zivojinovic, Velimir 'Bata' with 57 roles , Jeremy and Ron with 51 roles. The others are shown in figure 5.

Zivojinovic, Velimir 'Bata'	Jeremy, Ron
57	51
Doll, Dora	Dobtcheff, Vernon
47	47
Berléand, François	Galabru, Michel
42	42
North, Peter (I)	Renucci, Robin
42	42
Kapoor, Shakti (I)	Milinkovic, Predrag
41	41

FIGURE 5. Top ten most common actors.

Sam Neill and Martin Sheen were the top two connected actors. We looked at which actors have the largest betweeness. As shown in figure 6 we see the top ten connected actors.

If we look at two actors: Daniel Olbrychski and Sonia Braga we see one of the shortest path between them is from Daniel Olbrychski to Marie-Christine to Daniel Stern to Sonia Braga. So path had to go through two other actors. One would expect this for the top connected actors.

```
   Neill, Sam         Sheen, Martin     Dobtcheff, Vernon
     410102.4              299061.2              286767.6
Braga, Sonia Halmer, Günther Maria    Olbrychski, Daniel
     243651.4              232576.5              210122.9
  Jeremy, Ron        von Sydow, Max       Cheung, George
     209709.3              202576.6              190326.9
Kingsley, Ben
     187748.6
```

FIGURE 6. Top ten most connected actors.

```
#question 3
#order of the most common actor
o<-order(as.numeric(abs(nroles)),decreasing=T)
((nroles)[o[1:10]])
#those with the top betweeness
con_art = sort(betweenness(actnet),decreasing=T)[1:10]
con_art_names = names(con_art)
all_shortest_paths(actnet, from="Neill, Sam", to="Cheung, George")
for (i in 1:length(con_art)) {
  for (j in 1:length(con_art)) {
    if(con_art_names[i] != con_art_names[j] )

    {
      short_path = (get.shortest.paths(actnet, from=con_art_names[i], to=con_art_names[
    j]))
      print(con_art_names[i])
      print(con_art_names[j])
      print(short_path$vpath)

    }
  }
}
```

LISTING 3. R source code implemented for the homework.

Question 4. Find pairwise actor-cast association rules with at least 0.01 percent support and 10 percent confidence. Describe what you find.

The figure 7 shows the association of actors to Martin Sheen. For instance Sydney Penny and Martin Sheen; how likely am I to find Martin Sheen in a movie given Sydney Penny is in that same movie.

```

#question4
artrules <- apriori(casttrans,
                    parameter=list(support=1e-04, confidence=.1, maxlen=2))

## extract the rules as strings, then change into an edge matrix
## print pairs at each step to see what I've done

pairs <- labels(artrules)

pairs <- gsub("\\{|\\}","",pairs)

pairs <- strsplit(pairs," => ")

```

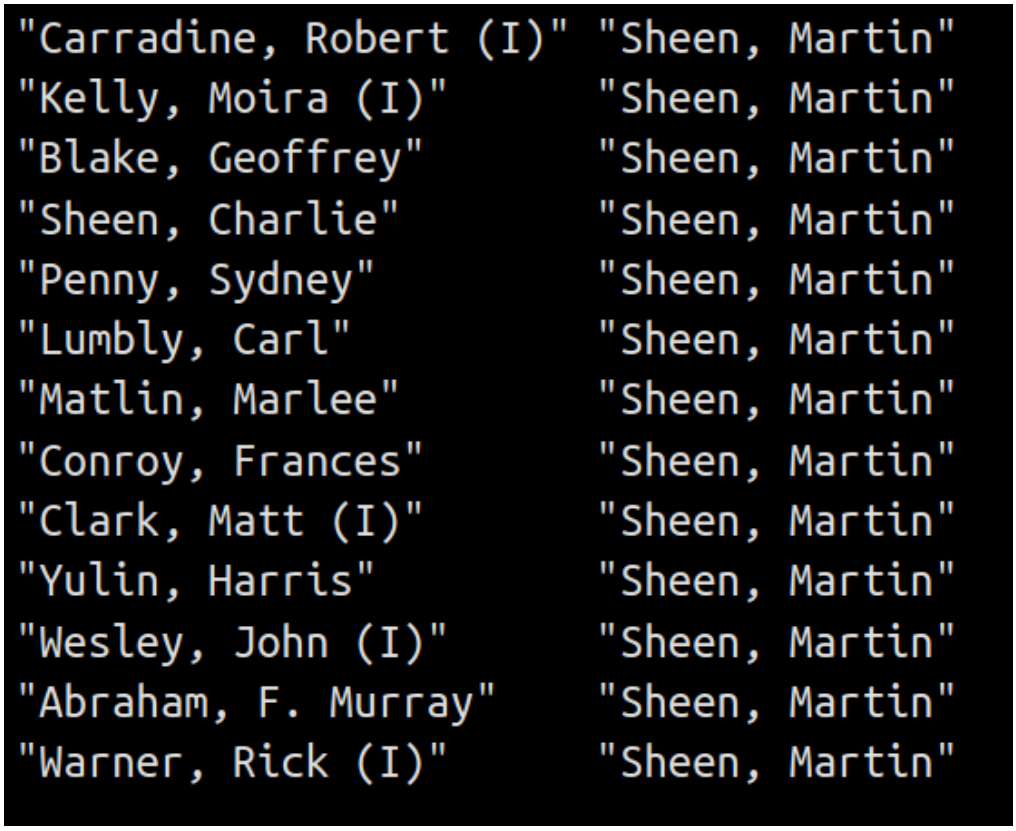

FIGURE 7. Association between other actors and Martin Sheen.

```
15
16
17 pairs <- do.call(rbind,pairs)
18
19 pairs <- pairs[pairs[,1]!="",]
20 left = which(pairs[,1] == "Sheen, Martin")
21 right = which(pairs[,2] == "Sheen, Martin")
22 pairs[left,]
23 pairs[right,]
```

LISTING 4. R source code implemented for the homework.

Question 5. Let's zoom in at "Sheen, Martin". Build a regression model to predict his presence in a production from other actors' presences. Who are the top 5 actors that increase Martin Sheen's chance in a production? Compare your finding with the one from association rules?

What we see are actors with an association with Martin Sheen. The results of our logistic regression model shows the actors with the top five coefficients. Look for instance to Sydney Penny who has a coefficient of 4.32. We see that exp(4.32) is $exp(4.32)$, which means that for every unit increase in x we get $exp(4.32)$ times higher odds. Interestingly enough we have multiple matches of names in question 4 and question 5. All the names that appear in figure 8 appear in figure 7 as it should. So we get agreement with both models.

```
1
2 #question5
3 library(gamlr)
4 v <- actmat[,"Sheen, Martin"]
5 X <- actmat[,colnames(actmat)!="Sheen, Martin"]
6 log_model = gamlr(X, v, family="binomial")
```

intercept	Carradine, Robert (I)	Penny, Sydney
-7.489679	4.929609	4.324045
Blake, Geoffrey	Abraham, F. Murray	Matlin, Marlee
4.183723	4.171332	4.097746

FIGURE 8. The top names with highest coefficients in our model.

```
summary(log_model)
# glm(v ~ actmat"[,-Sheen, "Martin], data=actmat)


o<-order(as.numeric(abs(coef(log_model))),decreasing=T)
(coef(log_model)[o[1:6],])caption
```