Network Analytics
Individual Assignment 2

February 8, 2022

INSTRUCTIONS

- This is an individual assignment.
- Submit your answer digitally as two files through Moodle:
  - An R markdown file (extension **Rmd**). Use the template provided to you and provide your answers (both code and text) below each question.
  - An **HTML** file "knitted" by RStudio including all the results and plots. More details on how to create these files will be provided in class on week 3.
- Follow the Style Guide (available on Moodle). You can be penalized on up to 20% in each question for which you do not follow the Style Guide.
- Questions regarding the assignment should be posted exclusively on the respective discussion forum on Moodle.

Deadline: Monday, February 21 at 23:59.

- Late submissions are not allowed

Warning: The detection of any form of plagiarism in your work means the assignment will be graded with ZERO points.

Movie Networks

We are interested in assessing what are the most important movies in the decade 2010-2019. We will use different strategies to do so. First, we will load and prepare the data.

Load and prepare the data

The first step is to load and prepare the movie data. The following instructions perform some routine data preparation operations. Each set of instructions is preceded by a comment explaining the procedure. Run the code below and try to understand each line of code as you might need to perform some changes.

```
library(data.table)     # Run once per session
library(ggplot2)        # Run once per session

# Load data from file 20200120-imdb_movie_actor.csv (do not forget to
# change your working directory to the folder containing the
# 20200120-imdb_movie_actor.csv file)
dt.movie.actor <- fread("20200120-imdb_movie_actor.csv")

# Count in how many movies each actor has participated and how many
# principal actor each movie has
dt.movie.actor[, n_movies := .N, by=actor]
dt.movie.actor[, n_actors := .N, by=list(movie, year)]

# Remove entries in which actors have no name
dt.movie.actor <- dt.movie.actor[!(actor == "")]

# Save dt.movie.actor. Next time you can simply call the load function (below)
save(dt.movie.actor, file="imdb_movie_actor.RData")
```

```
data.table 1.14.1 IN DEVELOPMENT built 2021-05-18 23:25:31 UTC; root using 6 threads (see ?getDTthreads).  Latest news: r-datata
**********
This development version of data.table was built more than 4 weeks ago. Please update: data.table::update.dev.pkg()
**********
Error in fread("20200120-imdb_movie_actor.csv") :
  File ’20200120-imdb_movie_actor.csv’ does not exist or is non-readable. getwd()==’/Users/rbelo/Dropbox (Erasmus Universiteit R
Error: object ’dt.movie.actor’ not found
Error: object ’dt.movie.actor’ not found
Error: object ’dt.movie.actor’ not found
Error in save(dt.movie.actor, file = "imdb_movie_actor.RData") :
  object ‘dt.movie.actor’ not found
```

Load the data that you prepared using the instructions below. As mentioned in the comments, you can start from this line if you have previously saved these data.

```
# Load previously saved dt.movie.actor. You can
# start in this line if you have previously saved these data.
load("imdb_movie_actor.RData")
```

```
Error in readChar(con, 5L, useBytes = TRUE) : cannot open the connection
In addition: Warning message:
In readChar(con, 5L, useBytes = TRUE) :
  cannot open compressed file ’imdb_movie_actor.RData’, probable reason ’No such file or directory’
```

Questions (`data.table`) `[7 points]`

This set of questions require that you know how to manipulate a `data.table`. Answer each of the following questions below. Include all the code you created/used in your answer.

1. What is the total amount of movies in the `dt.movie.actor` dataset? `[1 point]`
2. List the actors from the movie `"Fight Club (1999)"`. List the actors from the movie `"Se7en (1995)"`. `[1 point]`
3. Which actors participated on both movies? Hint: The function `intersect` calculates the intersection of two sets. `[1 point]`
4. In which movies did Brad Pitt (b.1963) and George Clooney (b.1961) star together? `[1 point]`
5. Create a table that shows the number of movies released per year. This table should include three columns: `year`, `n_movies`, and `csum_n_movies`. The first column should contain the year, the second the number of movies in that year, and the third, the number of movies released since the first year in the data and up to the year in that line. Tip: Use the function `cumsum` and check if the amount in the last year is the same as the total number of movies in question 1. `[1 point]`
6. Which actor/actress has starred in the most movies across all data? After (and including) 2000, which year has the most movie participations by a single actor/actress? Who is that actor/actress? What do these two actors/actresses have in common? `[1 point]`
7. Consider only the 10% most popular movies (by votes) in the decade 2010-2019. List the top 10 actors that starred in the most movies in the decade. Which year(s) has/have the most movie participations by a single actor? Hint: you can use the function `quantile` to find how many votes does the movie in percentile 90 have. `[1 point]`

Questions (`ggplot2`) `[3 points]`

1. Plot a histogram with the number of movies per year. Which patterns do you observe? Is there anything strange? `[1 point]`
2. Plot a histogram that represents the distribution of number of IMDb votes per movie. The x-axis should represent the number of votes and the y-axis should represent how many movies have x number of votes. Which patterns do you observe? `[1 point]`
3. Plot a histogram that represents the distribution of the number of actors per movie. The x-axis should represent the number of actors and the y-axis should represent how many movies have x number of actors. Describe your findings. `[1 point]`

Questions (igraph) [10 points]

1. From this question onwards, and until the end of the assignment, focus only on the actors that participated on the top 50 most popular movies from the 2010-2019 decade (by number of votes). Load the `igraph` package and create a bipartite graph in which the edges correspond to actors' participation in movies. How many movie participations exist? `[1 point]`

2. Create a graph in which two movies are connected to each other if they have at least one actor in common. Calculate the degree centrality for each of the movies, and remove movies with no connections to other movies. Hint: the function `induced.subgraph` allows the creation of graphs with only a subset of the vertices. Calculate the following additional centrality measures for each of these movies: `[2 points]`

   - Closeness centrality
   - Betweenness centrality
   - Eigenvector centrality

3. For each centrality measure, list the top 20 movies with highest centrality. How do you interpret the outcomes? `[2 points]`

4. Calculate the average clustering coefficient for the movies network. `[1 point]`

5. Choose one movie you like and plot the movie, their direct neighbors and the links among them. What is the clustering coefficient of this movie? Which is the actor with most participations among these (neighbor) movies, but not having participated in the movie itself? `[2 points]`

6. Plot the degree distribution of the movies. How do you compare them with the degree distribution of a random graph? What can be plausible explanations for the observed differences? `[2 points]`