vlado

Star Wars networks

Feb 9-10, 2025

Kaggle

The Kaggle [https://www.kaggle.com/datasets/mexwell/star-wars-network/versions/1] dataset comprises two CSV files:

`starwars-characters.csv`: Contains a list of characters with the following columns:

- number: Unique identifier for each character.
- name: The name of the character.
- scenes: Number of scenes in which the character appears.

`starwars-links.csv`: Captures the interactions between characters with the following columns:

- character1: Name or ID of the first character.
- character2: Name or ID of the second character.
- scenes: Number of scenes in which both characters appear together.

The creation of the Kaggle Star Wars data is described at GitHub/chatox [https://github.com/chatox/networks-science-course/tree/master/practicum/data/starwars] where we learn that it is based on the data collected by Evelina Gabasova [https://github.com/evelinag/star-wars-network-data].

Evelina Gabasova. (2016). Star Wars social network (Version 1.0.1) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1411479 [http://doi.org/10.5281/zenodo.1411479]

Kaggle -> Pajek

```
> wdir <- "C:/Users/vlado/DL/data/kaggle/StarWars"
> setwd(wdir)
> source("https://raw.githubusercontent.com/bavla/Rnet/master/R/Pajek.R")
> V <- read.csv("starwars-characters.csv")
> head(V)
  number        name scenes
1      0 DARTH VADER    190
2      1       R2-D2    171
3      2   CHEWBACCA    145
4      3        BB-8     40
5      4     QUI-GON     62
6      5 NUTE GUNRAY     25
> L <- read.csv("starwars-links.csv")
> head(L)
  character1 character2 scenes
1          0          1     32
2          2          0      2
3          0         20      5
4          0          4     22
5          0         18     41
6          0         21      2
> dim(V)
[1] 111   3
> dim(L)
[1] 444   3
> n <- nrow(V); m <- nrow(L)
> M <- matrix(NA,nrow=n,ncol=n)
> rownames(M) <- colnames(M) <- V$name
> diag(M) <- V$scenes
> for(r in 1:m){i <-L$character1[r]+1; j <- L$character2[r]+1; v <- L$scenes[r];
+ if(is.na(M[i,j])) M[i,j] <- v else cat(r,i,j,"\n") }
> M[1:5,1:5]
            DARTH VADER R2-D2 CHEWBACCA BB-8 QUI-GON
DARTH VADER         190    32        NA   NA      22
R2-D2                NA   171        NA   NA      NA
CHEWBACCA             2    17       145   NA      NA
BB-8                 NA     2         8   40      NA
QUI-GON              NA    13        NA   NA      62
> matrix2net(M,Net="StarWars.net",nolink=NA)
```

The obtained network `StarWars.net` is directed. I transformed it into an undirected network using Pajek and saved it as `StarWarsE.net`.

Evelina

Evelina's data are more detailed - they determine a temporal multi-relational weighted network. 7 episodes determine the time 1-7; time = 8 corresponds to all 7 episodes merged. There are 3 relations: 1-mentions, 2-interactions, and 3-interactions-allCharacters.

Evelina -> Pajek

```
> wdir <- "C:/Users/vlado/DL/data/kaggle/StarWars"
> setwd(wdir)
> library(jsonlite)
> F <- c("starwars-episode-*-mentions.json",
+        "starwars-episode-*-interactions.json",
```

```
+        "starwars-episode-*-interactions-allCharacters.json")
> T <- c("m","i","a")
> SW <- fromJSON("starwars-full-mentions.json")
> N <- sort(SW$nodes$name); n <- length(N)
> I <- order(SW$nodes$name)
> nodes <- data.frame(name=N,color=SW$nodes$colour[I])
> links <- NULL
> for(r in 1:3){
+   for(e in 1:8){
+     if(e==8) {fjson <- gsub("episode-\\*","full",F[r]); col <- T[r]} else
+       {fjson <- gsub("\\*",as.character(e),F[r]); col <- paste0(T[r],e)}
+     SW <- fromJSON(fjson)
+     p <- match(SW$nodes$name,N); V <- rep(0,n); V[p] <- SW$nodes$value
+     nodes[[col]] <- V; m <- nrow(SW$links)
+     u <- as.integer(factor(SW$nodes$name[SW$links$source+1],levels=N))
+     v <- as.integer(factor(SW$nodes$name[SW$links$target+1],levels=N))
+     w <- SW$links$value
+     L <- data.frame(n1=u,n2=v,w=w,t=rep(e,m),r=rep(r,m))
+     links <- rbind(links,L)
+   }
+ }
>
> SW <- list(nodes=nodes,links=links)
```

We combined nodes and links into a network SW. We save it in R format for potential later use.

```
> str(SW)
List of 2
 $ nodes:'data.frame':  113 obs. of  26 variables:
  ..$ name : chr [1:113] "ADMIRAL ACKBAR" "ADMIRAL STATURA" "ANAKIN" "BAIL ORGANA" ...
  ..$ color: chr [1:113] "#808080" "#808080" "#ce3b59" "#808080" ...
  ..$ m1   : num [1:113] 0 0 62 3 0 0 0 0 0 0 ...
  ..$ m2   : num [1:113] 0 0 62 4 0 0 7 0 0 10 ...
  ..$ m3   : num [1:113] 0 0 105 28 0 0 2 0 0 0 ...
  ..$ m4   : num [1:113] 0 0 0 0 0 0 6 0 34 0 ...
  ..$ m5   : num [1:113] 0 0 0 0 0 0 0 0 0 10 ...
  ..$ m6   : num [1:113] 11 0 6 0 0 0 0 7 0 5 ...
  ..$ m7   : num [1:113] 5 3 0 0 4 59 0 0 0 0 ...
  ..$ m    : num [1:113] 15 3 232 33 4 59 13 7 34 23 ...
  ..$ i1   : num [1:113] 0 0 40 2 0 0 0 0 0 0 ...
  ..$ i2   : num [1:113] 0 0 39 3 0 0 2 0 0 6 ...
  ..$ i3   : num [1:113] 0 0 53 20 0 0 0 0 0 0 ...
  ..$ i4   : num [1:113] 0 0 0 0 0 0 5 0 18 0 ...
  ..$ i5   : num [1:113] 0 0 0 0 0 0 0 0 0 5 ...
  ..$ i6   : num [1:113] 9 0 2 0 0 0 0 3 0 0 ...
  ..$ i7   : num [1:113] 4 3 0 0 4 2 0 0 0 0 ...
  ..$ i    : num [1:113] 12 3 131 23 4 2 6 3 18 10 ...
  ..$ a1   : num [1:113] 0 0 41 3 0 0 0 0 0 0 ...
  ..$ a2   : num [1:113] 0 0 40 4 0 0 3 0 0 7 ...
  ..$ a3   : num [1:113] 0 0 54 21 0 0 0 0 0 0 ...
  ..$ a4   : num [1:113] 0 0 0 0 0 0 6 0 19 0 ...
  ..$ a5   : num [1:113] 0 0 0 0 0 0 0 0 0 6 ...
  ..$ a6   : num [1:113] 10 0 3 0 0 0 0 4 0 0 ...
  ..$ a7   : num [1:113] 5 4 0 0 5 40 0 0 0 0 ...
  ..$ a    : num [1:113] 13 4 132 24 5 40 7 4 19 11 ...
 $ links:'data.frame':  3827 obs. of  5 variables:
  ..$ n1: int [1:3827] 69 77 70 70 70 81 81 102 69 32 ...
  ..$ n2: int [1:3827] 81 102 81 110 102 110 102 110 102 69 ...
  ..$ w : int [1:3827] 3 1 46 7 1 6 1 1 2 1 ...
  ..$ t : int [1:3827] 1 1 1 1 1 1 1 1 1 1 ...
  ..$ r : int [1:3827] 1 1 1 1 1 1 1 1 1 1 ...
> saveRDS(SW,"StarWars.rds")
> # SW <- readRDS("StarWars.rds")
```

Finally, we export the network and network partitions in Pajek format.

```
> source("https://raw.githubusercontent.com/bavla/Rnet/master/R/Pajek.R")
> u <- links$n1; v <- links$n2; levels(u) <- N; levels(v) <- N
> uvFac2net(u,v,w=links$w,r=links$r,t=links$t,Net="StarWarsC.net")
> vecnom2clu(nodes$color,Clu="SWcolor.clu")
> CN <- colnames(nodes)[-c(1,2)]
> for(nam in CN) vector2clu(nodes[[nam]],Clu=paste0("SW",nam,".clu"))
```

I combined in Pajek all created Pajek files (network and partitions) into a single Pajek project file `StarWars.paj`. I manually changed Arcs → Edges, added some metadata, and made some cleaning.

In Pajek we can now extract subnetworks of our interests.
