Assignment Nº1

851-0252-06L Introduction to Social Networks: Theory, Methods and Applications FS2023

Daniel Repérant
Julia Salustowicz
Maria Eugenia Gil Pallares
Zhengxu Li

Task 1: Describe and plot a social network

(a) load the affective network (2400 affective w1.csv) and the gender data (2400 sex.csv)

We load the data with function read.csv(), and transform it into a matrix, it is worth to mention that we need to take care of the first column of the csv file as we just use the id of student as header but do not include them into the adjacency matrix.

```
setwd("The_path_of_work")
#we load the data and transform it into a matrix
affective = read.csv("2400_affective_w1.csv")
# we don't want the first column into our adjacency matrix as it is just the id
affective=as.matrix(affective[,-1])
#we put id as row and column names
rownames(affective)=colnames(affective)
#another dataset of gender, we apply same way of reading data as previously
sex = read.csv("2400_sex.csv")
id = sex[,1]
sex=as.matrix(sex[,-1])
rownames(sex)=id
```

(b) recode the affective network of wave 1 into a friendship network. (Hint: the value +2 stands for friendship ties – see the data description for details)

We suppose that there is only friendship if there is a 2 in the matrix, as a value 2 indicates best affection and 2 stands for friendship relationship as the statement claims, it also helps as a preparation for the coming subtask using trust network, as friends are more prone to share secrets.

```
friendship=affective #we copy the affective network to a new variable to change values there
friendship[!friendship==2]=0 # all the relationships with value different than 2 will not be considered
friendship[friendship==2]=1 #we use only friendship ties
```

(c) calculate basic network descriptives for this friendship network

• network size (i.e., number of nodes), density, average degree, reciprocity ratio, gender composition in class, count of same gender ties (i.e., both nodes have the same gender)

network size = 27

```
ncol(friendship)
```

density = 0.1880342

```
gden(friendship)
```

average degree = 4.888889

```
sum(friendship, na.rm = T)/dim(friendship)[1]   # how many ties each person has on average
```

reciprocity ratio = 0.3608247

```
grecip(friendship, measure = "dyadic.nonnull")
```

gender composition in class : 11 males and 16 females

```
table(sex)
```

count of same gender ties = 108(directed ties, no matter if a tie is bidirectional/mutual, we count the edge even is unidirectional, 36 for male and 72 for female, 108 in total), or
count of same gender ties = 34(if we only count mutual ties, there are 22 mutual ties for female and 12 mutual ties for male)

```
index_men=which(sex==1) #we find index of men and women to locate them
index_women=which(sex==2)
friendship_women=friendship[index_women,index_women]  #we build 2 submatrix with only women and men
friendship_men=friendship[index_men,index_men]
men.igraph <- graph_from_adjacency_matrix(friendship_men, mode = "directed")
women.igraph <- graph_from_adjacency_matrix(friendship_women, mode = "directed")
gsize(men.igraph)
gsize(women.igraph)
sum(which_mutual(men.igraph))/2 #divide by two as we count two mutual ties as 1 connection
sum(which_mutual(women.igraph))/2
```

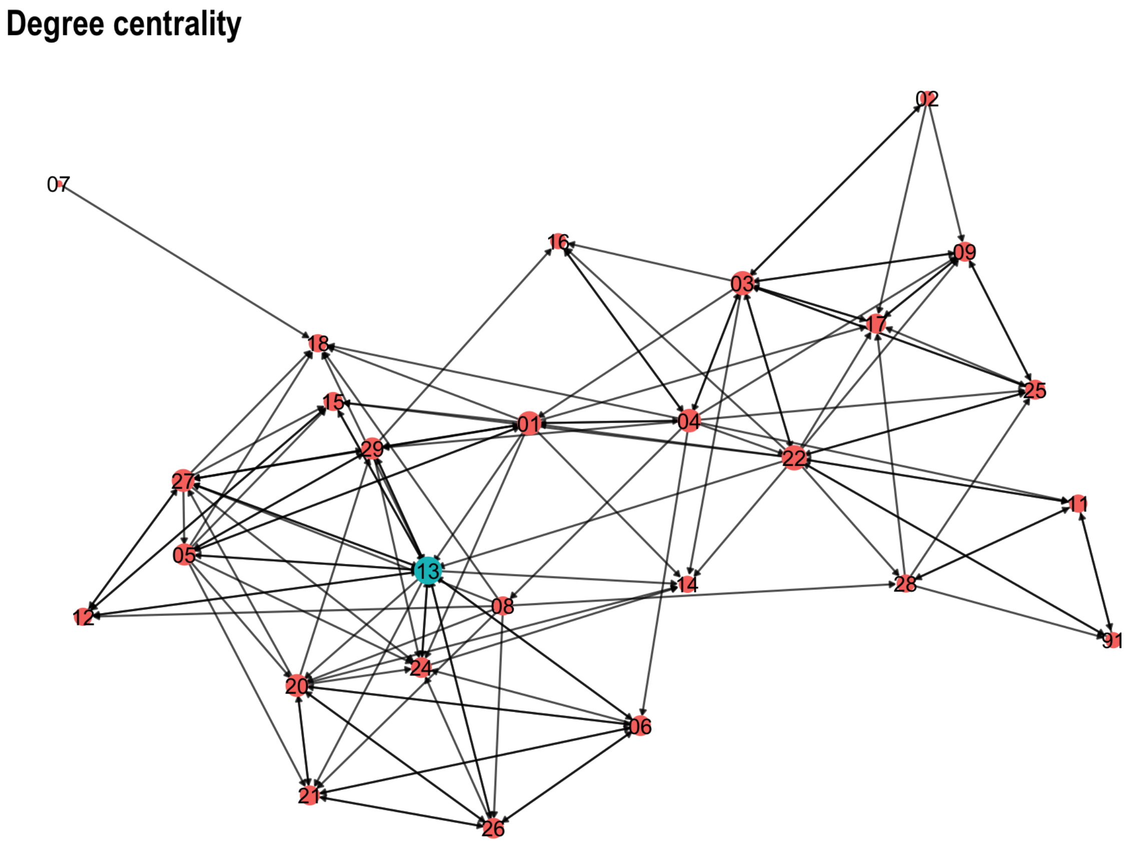
• plus one other measure of your choice
We calculate degree centrality as our choice of network descriptive. The degree centrality of the nodes are: 15 4 15 14 12 10 1 8 9 7 7 22 6 8 5 9 7 13 9 17 10 9 10 13 7 12 5
Where the maximum is the node 2413 that has 22 degree centrality.

```
degree <- sna::degree(friendship, cmode = 'freeman')
# freeman specifies the total of incoming and outgoing ties
max(degree) # using max(), we see the highest value
degree # the sequence of degree centrality of the nodes
```

```
which(degree == max(degree)) # the 12th person got highest degree which is "2413"
```

• briefly interpret the measures (where sensible)

We used the function degree() from library sna, with parameter mode = “freeman”,which indicates that we take into account the sum of in-degree and out-degree(i.e. number of incident edges it has). The higher the degree, the more central the node is. This can be an effective measure, since many nodes with high degrees also have high centrality by other measures.



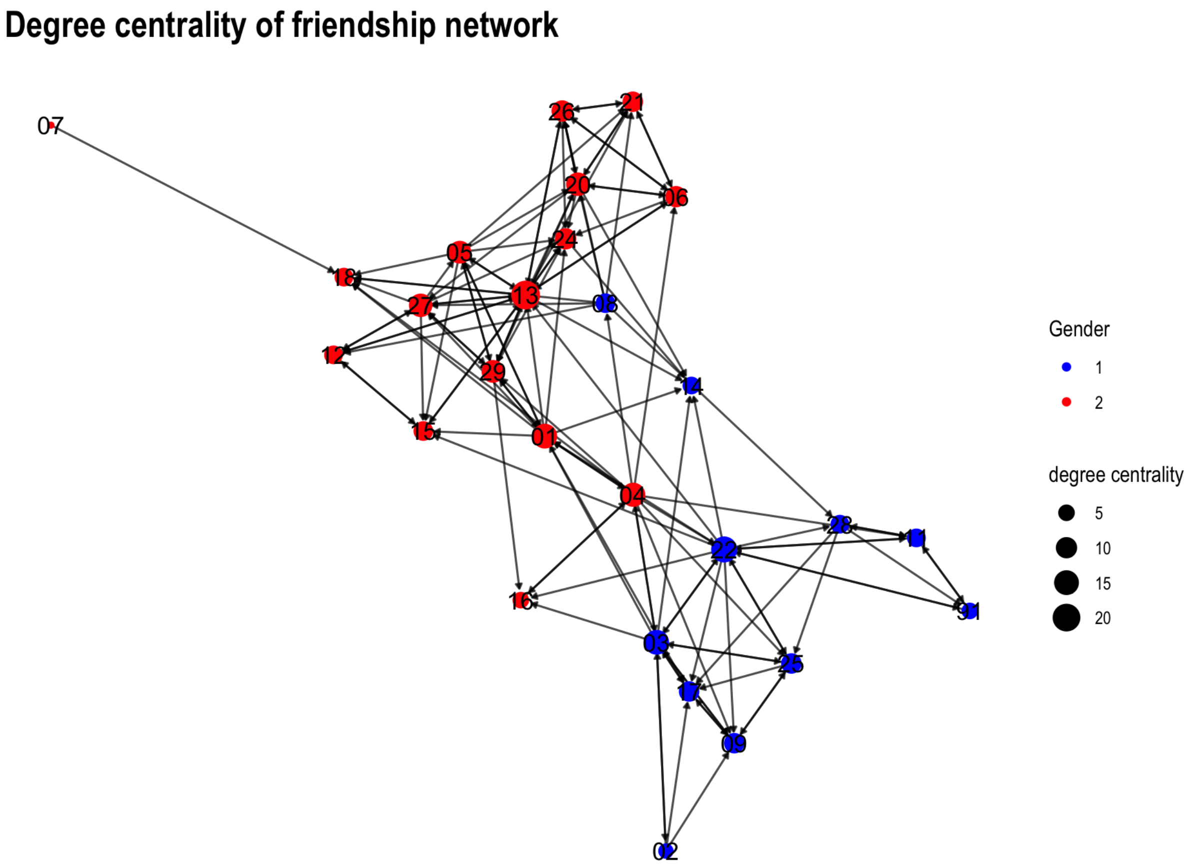
Here we visualize the degree centrality of our nodes, where the size of the node is proportional to the degree centrality, we can observe that the node 2413 with the highest degree centrality is coloured with blue, which is the most “popular” node in our network, and another nodes with lower degree centrality are colored in red.

(d) plot the friendship network

• the plot has to be informative
• color the nodes according to the gender of the person
• the node size should be proportional to a centrality measure (of your choice).

Here we plot the friendship network as required, as it can be read from the legend, we color the male node with blue and female nodes with red, and the size of our vertices are proportional to the degree centrality.

It can be observed in our graph that the group of male students are grouped in one cluster and the group of female students are grouped in another cluster. The center nodes of each cluster have larger size as there are more incident edges connected to these popular nodes(for instance, node 13 and node 22). Nevertheless, the peripheral nodes have smaller size as they are less "popular".

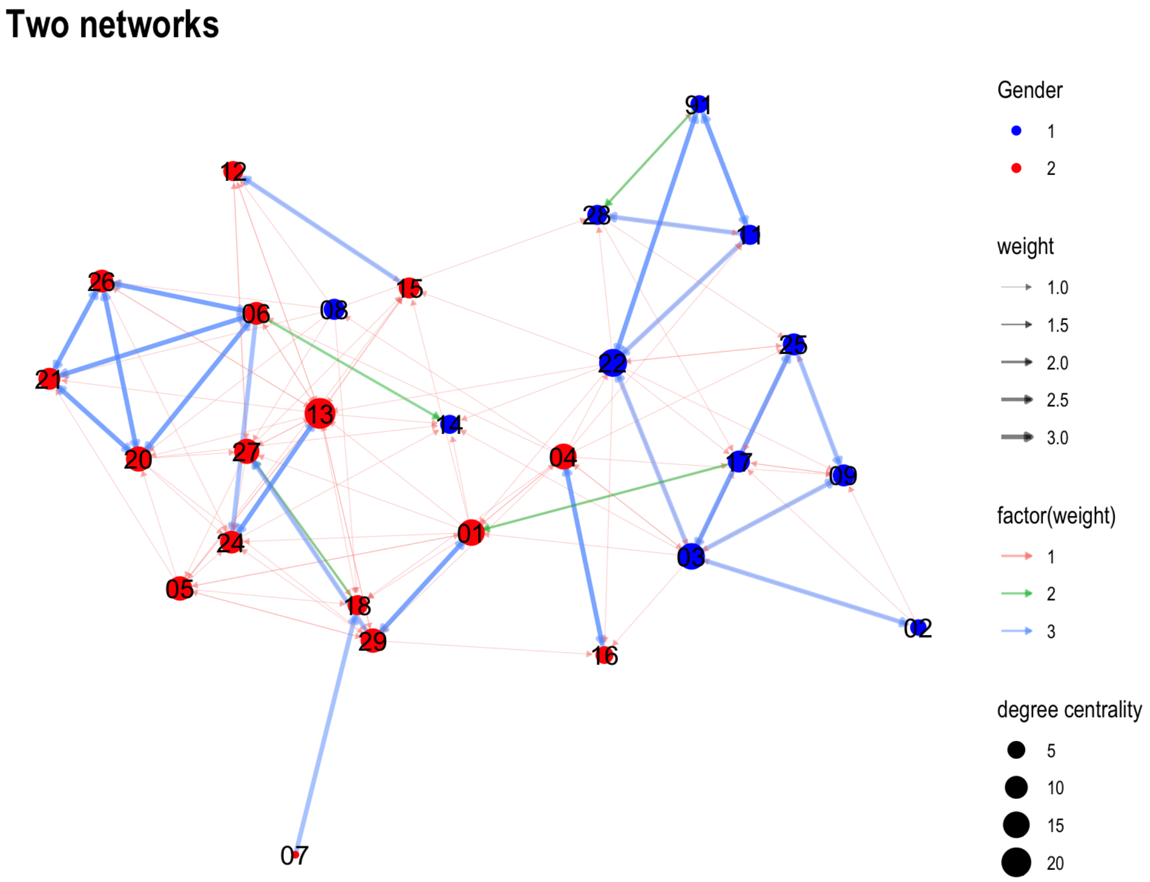


```
set.seed(53)
#create a layout
layout.graph <- create_layout(friendship.igraph,  layout = 'fr')
#we adjust the aesthetic of the edges
edges  <-  geom_edge_link(  alpha=0.7,  arrow=arrow(length  =  unit(1,  'mm'),  type  =
"closed"), end_cap = circle(2, 'mm'))
#we color the nodes according to the gender and make the size proportional to the
degree centrality
nodes <- geom_node_point(aes(colour = as.factor(sex), size = degree))
#we put some legends to facilitate understanding
plotlabs <- labs(colour = "Gender",size = "degree centrality")
```

```
#we draw the plot
degr.plot <-
  ggraph(layout.graph)+
  edges+
  nodes+
            geom_node_text(aes(label = substr(colnames(friendship),4,5),size = 10),nudge_x=0, nudge_y=0)+ #put the id of student on the node
  plotlabs+
  scale_color_manual(values = c("1" = "blue", "2" = "red")) + #blue-male, red-female
  theme_graph()+
  ggtitle("Degree centrality of friendship network")
degr.plot
```

(e) Now also include the trust network (2400 trust w1.csv) and plot them both in one network plot (i.e., on top of each other). (Hint: check the forum on moodle for an example)

We used the hint stated on the forum creating a weighted matrix net3 where edges with different values represent different types of combination.



To enhance the visual clarity of our graph, we have employed a color-coding scheme for different types of edges. Specifically, edges with a value of 1 (i.e., those that only appear in network 1) are represented in orange, while those with a value of 2 (exclusive to network 2) are shown in green. Edges with a value of 3 (that appear in both networks) are depicted in blue. To further distinguish between these three types of connections, we have adjusted the width of the edges accordingly, although it is worth noting that there are only three distinct widths. By incorporating both color and width scaling based on edge weight, we can more easily comprehend the relationships between nodes in the network.

```
trust= read.csv("2400_trust_w1.csv")
trust=as.matrix(trust[,-1])
rownames(trust)=colnames(trust)

net1 <- friendship
net2 <- trust
# a new matrix of the right size, prefilled with ties in net1
net3 <- net1
# where ties exist in the second network, we give a new value
net3[net2 == 1] <- 2
# where ties exist in both networks, we replace with a new, unique value
net3[net1 == 1 & net2 == 1] <- 3
# now you have 4 unique values in net3:
table(net3)

# 0 = no ties,
# 1 = tie in network 1,
# 2 = tie in network 2,
# 3 = tie in both networks

set.seed(50)# set a random seed to make it reproducible
g=graph_from_adjacency_matrix(net3,mode="directed",weighted=TRUE)
#now we try a different algorithm of layout
layout.graph <- create_layout(g, layout = 'stress')
# we color the edges and make the width proportional to the weight in order to make better
visualization
edges <- geom_edge_link2(alpha = .5,
                         arrow = arrow(length = unit(1, 'mm'),
                                       type = "closed"),
                         end_cap = circle(2, 'mm'),aes(width = weight,color=factor(weight)))
#nodes are colored according to gender and size proportional to the degree
nodes <- geom_node_point(aes(colour = as.factor(sex), size = degree))
plotlabs <- labs(colour = "Gender",size = "degree centrality")

# we draw the plot
degr.plot <-
  ggraph(layout.graph)+
  edges+
  nodes+
     geom_node_text(aes(label = substr(colnames(friendship),4,5),size = 10),nudge_x=0,
nudge_y=0)+ # add id
  plotlabs+
  scale_color_manual(values = c("1" = "blue", "2" = "red")) +
  scale_edge_width(range = c(0.1, 1))+ #scale the width of the edge to make the type of ties
more relevant
  theme_graph()+
  ggtitle("Two networks")

degr.plot
```

(f) How large is the overlap between the two networks? (i.e., how many ties between two people are present in both networks)

34 edges

We count the number of value 3 in the weighted matrix, which corresponds to the number of edges that are present in network 1 and network 2, we got 34.

```
sum(net3 == 3, na.rm = TRUE)
```

(g) In a short paragraph (max. 250 words), describe what you see in the network plot and comment on the overlap between the two networks.

Upon examining the network plot, we can discern that there are considerably more orange ties (98 edges that are only present in network 1) compared to green ties (edges that are exclusive to network 2). This leads us to infer that while many students may consider others as their friends, they may not necessarily be willing to confide in them. Conversely, the existence of only 4 green ties (edges that only appear in network 2) indicates that those whom you are willing to share secrets with are likely to be considered as your friends. In the previous exercise, we identified 34 overlapping edges between the two networks. Coupled with the fact that there are 98 edges in network 1, we can conclude at a global level that approximately one-third of the friends (34/98) are willing to share secrets. Furthermore, it is evident that there exist three minor groups with a significant level of trust (indicated by blue edges). Notably, there are no instances of high trust connections between male and female clusters. Despite being popular nodes with high centrality degrees in the friendship network, nodes 2413 and 2422 do not possess many high-trust connections. As a result, it is plausible to assume that an increase in popularity may lead to less focus on maintaining strong friendships, ultimately resulting in a decrease in the level of trust in the ties.

All answers to these questions need to be printed in the PDF file. We will not run your R-Scripts to correct the assignment – we only use it to recreate your thought process in case the answers are wrong.

Task 2:

a) **Compare the advantages and disadvantages of complete network data, snowball-sampled network data, and ego-centered network data. Describe at least one advantage and disadvantage for each type of data. (max. 400 words)**

Complete network data:

- Advantages: Complete network data sampling is easier to collect as the population and boundaries are known. Access to the sample population is easier and quicker, less costly, for example one could send out an e-mail to the whole sample group that is normally an organization or school or something alike.
- Disadvantages: Individuals that are part of the closed sampling group could falsify the data to their or others advantages. As the boundaries of the sample population are known, anonymity is weakened. Also, the filled out survey yield can be lower than usual, as the survey is big, the questions one should answer apply to all other group members, which make the cost of filling it out way higher.

Snowball-sampled network data:

- Advantages: Snowball-sampled network data provides a good random initial seed of study participants. The initial participants are sampled not in a domain with known boundaries, which makes them chosen “more randomly” than for example in a complete network survey. Acquisition of new participants is done by already involved participants, making the search for people of the target group easier. Also, if you manage to get two new participants out of one already involved, the dataset growth is exponential.
- Disadvantages: As the acquisition of new study participants relies heavily on initial participants, it may be difficult to get new participants in studies where information that is collected has a negative connotation or is illegal, e.g. the usage of drugs or the sharing of needles. Some participants may not want to involve others, so the incentives to get them to share the information must be adequately high.

Ego-centered network data:

- Advantages: Very detailed surveys are possible as the survey effort is concentrated around the ego. The ego has not as many questions to answer as for example in complete network data. Also, ego-centered datasets can highlight psychological or social beliefs of single individuals, whereby snowball-sampled or complete network data shows more general trends within a group or population.
- Disadvantages: Vulnerability against falsification of data to the advantage of the ego is pronounced, as the survey depends on the ego and a falsification would affect the whole network.

b) **Describe common and unique ethical challenges that these three types of network data have and how you would address them if you were to conduct an empirical study. List at least one common ethical challenge and one unique ethical challenge for each type of network data (total of 4 challenges). (max. 400 words)**

Common Challenge: Surveys are vulnerable to privacy and anonymity issues. Inside observers are more likely to guess which individual is represented in the network by analyzing the node's positions in the network. Or there could also be a data breach, which could lead to malicious third parties to exploit the data. Avoid this issue by (i) not showing the results to the participants and/or (ii) set in place good anonymization and security techniques like data perturbation or swapping.

Complete network data:

- unique ethical challenge: As the boundaries and conditions in which the target group is functioning in are known, asking weird / sensitive / insensitive or inadequate questions that are not normal for this group might let the individuals feel like they misbehaved or did something bad, or maybe they suddenly think they do not conform to the norms of this group they are part of. Avoid this by really considering if a question is adequate for the survey setting.

Snowball-sampled network data:

- unique ethical challenge: There is a unique risk regarding the acquisition of new participants: the contacts or potential new participants that have been nominated by already involved participants have not consented to being approached by the survey conductors. The privacy of such contact information must be held intact. This issue could be solved by asking the current participants to personally approach their acquaintances and ask if they want to participate in the survey before the survey conductors officially contact the new participants.

Ego-centered network data:

- unique ethical challenge: Since the individual is in the focus of the survey, the ego might get throughout the survey a feeling of being "interrogated". Also, as the survey's success seemingly relies on the answers of the ego, it might also feel pressured to answer any questions asked, maybe answering questions they would normally not answer. To avoid this issue, make sure to get the ego comfortable and provide clear possibilities to opt out of answering certain questions.

c) **Choose one of the three types of network data and come up with a research question that you (i) can and (ii) cannot address with this type of data. Indicate why. (Hint: A research question indicates what you want to find out in a particular piece of scientific work. A question that is presented to participants (e.g., 'who are your friends?') is not a research question.) (max. 400 words)**

For the snowball-sampled network data one could ask "What is the average chain length from producer to consumer of cannabis in New York?". This works, because the initial participants introduce only people of interest for the research. With enough large samples statistically one could find enough producers of cannabis to generate significant data.
A research question, which one could not address with this type of data, would be: "Does meeting with your work team more than twice a week decrease your productivity?". This data would not give a good statistical input. This research question could instead be used in a complete network data.