Final exam. 2nd part

Name: Reinier Mujica

Instructions: This part is worth 5.5 points (plus 1 bonus point for *(h)*). You must provide the Rmd and html files with your answers through the *Final Exam* task. Do not change the global options chunk, except, if you feel it necessary, the figures' dimensions (in particular, the resulting html file must contain the displayed code chunks). The grading will take into account the cleanness of the html file. When explanations or comments are asked for, please remember to provide them. But, again, a typical explanation/comment should be a couple of sentences long, and unnecessarily long answers will be penalized. You can provide your answers in Catalan, Spanish, English, or French.

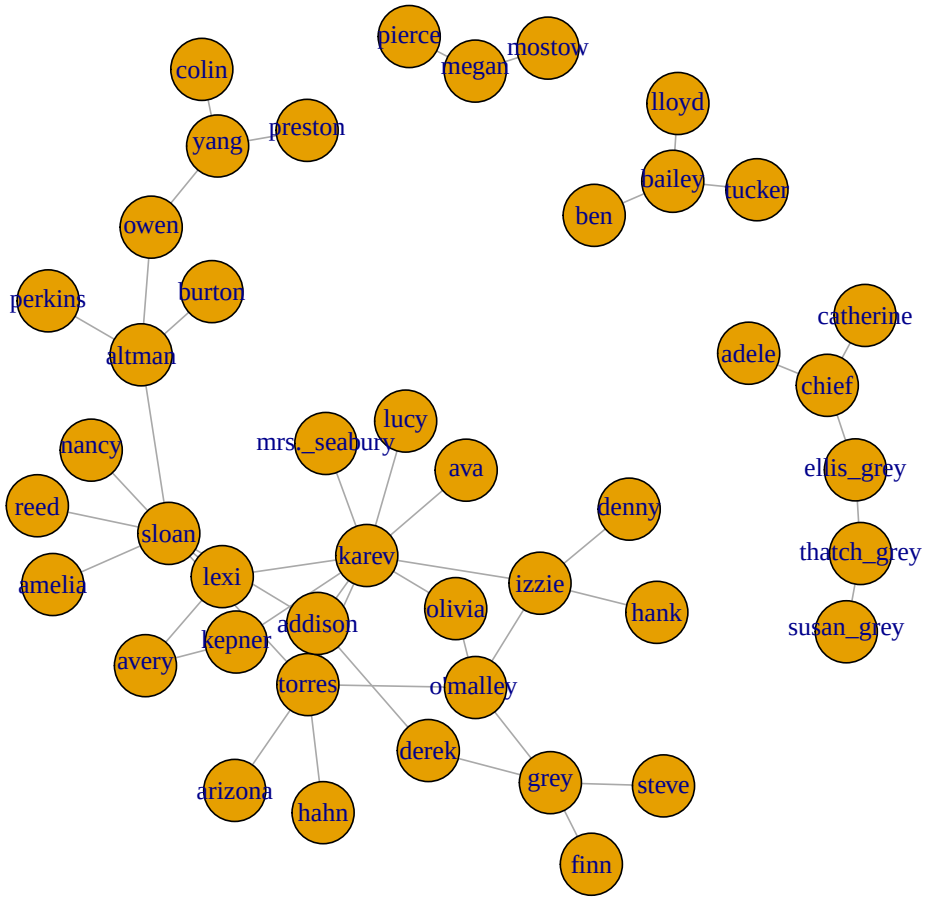
1) The files **GA_edges.txt** and **GA_nodes.txt** contain the links and nodes (with several attributes) of the sexual contact network among "Grey's Anatomy" characters in seasons 1-8. Define a network with these data frames, and make sure that its nodes have as attributes at least their names and (anatomic) sex.

Clean Up

```
rm(list=ls())
setwd('d:/MADM/redes sociales/FINAL/')
```

Define a network with these data frames, and make sure that its nodes have as attributes at least their names and (anatomic) sex.

```
ga_nodes <- read.table("GA_nodes.txt", header=TRUE, as.is=T)
ga_edges <- read.table("GA_edges.txt", header=TRUE, as.is=T)

ga <- graph_from_data_frame(d = ga_edges, directed = FALSE)

V(ga)$name = as.character(ga_nodes$name[match(V(ga)$name, rownames(ga_nodes))])
V(ga)$sex = as.character(ga_nodes$sex[match(V(ga)$name, ga_nodes$name)])

plot(ga)
```



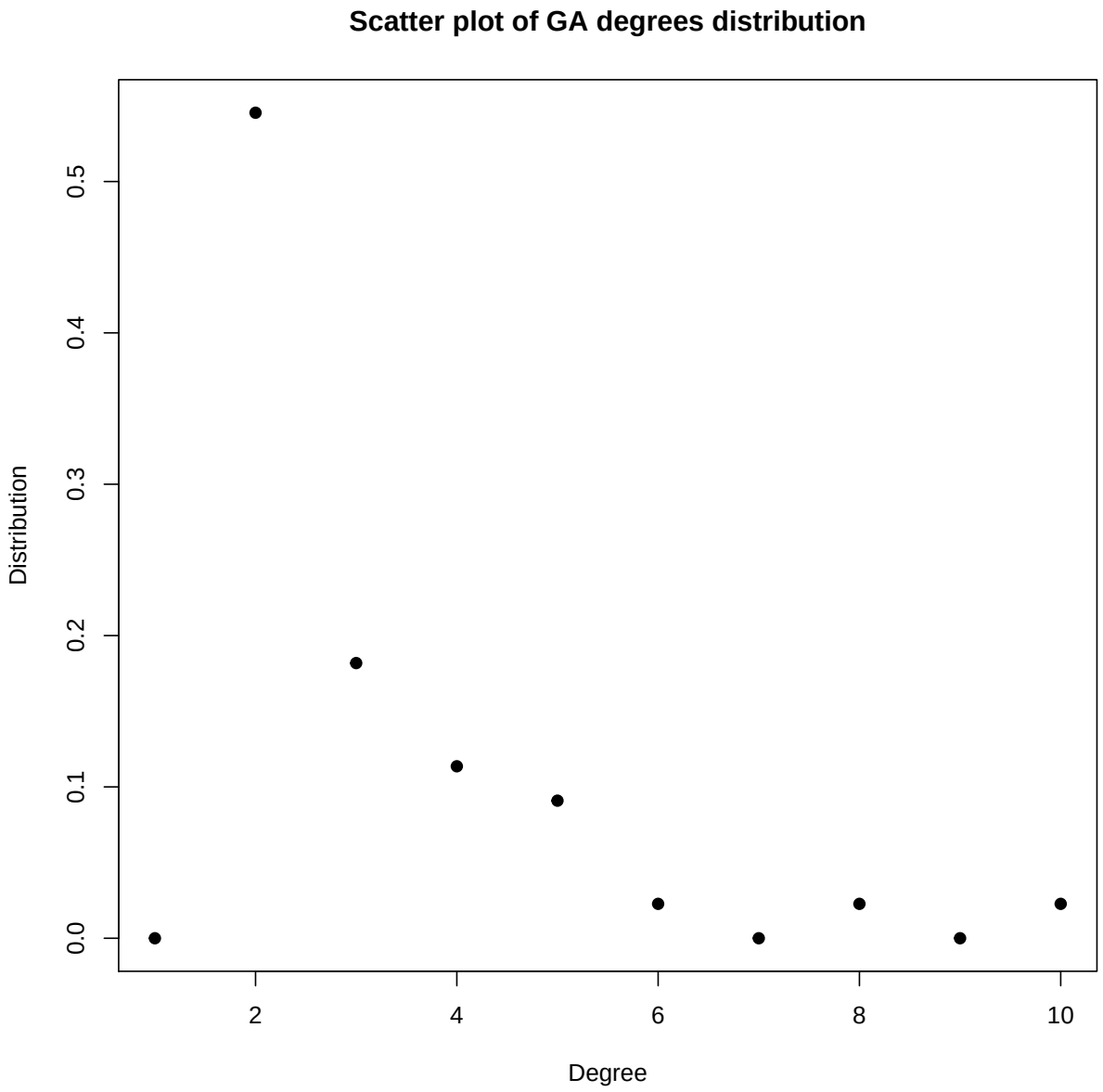
a) Provide a statistical summary of this network: order, number of nodes of each sex, size, density, number of connected components, **size of the largest connected component if there are more than one**, average degree, scatter plot of its degrees distribution, average distance, diameter, and average clustering coefficient. Plot it with the nodes labelled with the characters' names and differently colored with their sex.

```
ga_order <- gorder(ga)
ga_size <- gsize(ga)
ga_dens <- round(edge_density(ga),4)

ga_male = sum(V(ga)$sex == "M")
ga_female = sum(V(ga)$sex == "F")

ga_cc <- components(ga)$no
ga_cc_size_lg = components(ga)$csize[1]
```

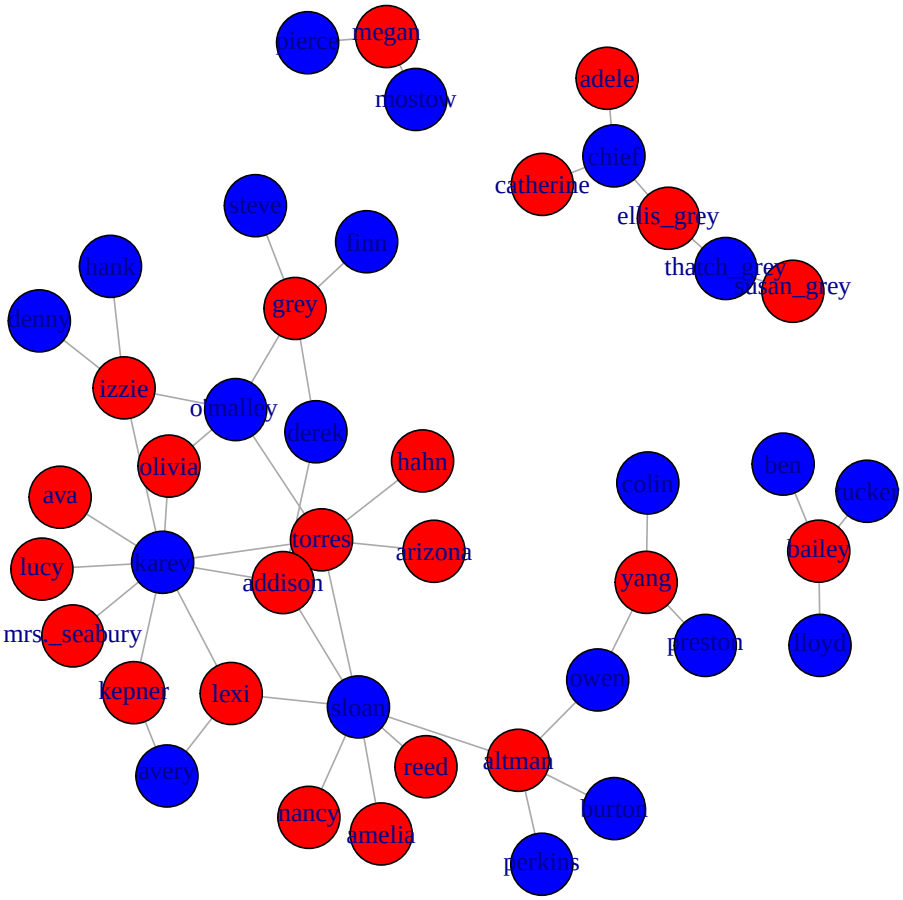
```
ga_avgdeg <- round(mean(degree(ga)), 0)

ga_diam <- diameter(ga)
ga_avgdist <- round(mean_distance(ga), 0)
ga_avg_clust_coef <- transitivity(ga, type="average")
```

- Order: 44
- Number of male nodes: 21
- Number of female nodes: 23
- Size: 46
- Density: 0.0486
- Number of connected components: 4
- Size of the largest connected component: 31
- Average degree: 2
- Average distance: 3
- Diameter: 8
- Average clustering coefficient: 0

```
ga_deg_dist = degree_distribution(ga)
plot(1:length(ga_deg_dist), ga_deg_dist, main="Scatter plot of GA degrees distribution", xlab="Degree",
```




```
V(ga)$color=V(ga)$sex
V(ga)$color=gsub("F","red",V(ga)$color) #Females will be red
V(ga)$color=gsub("M","blue",V(ga)$color) #Males will be blue
plot(ga)
```



a.2) Comment the meaning and implications (for this specific network) of the density, number of connected components and size of the largest, average degree, average distance, and average clustering coefficient values obtained.

La densidad de la red es muy baja 0.0486, esto implica que no todas los personajes tuvieron relaciones entre todos ellos, el número de componentes conexos es 4, es muy bajo y el primer componente tiene un tamaño de 31 que es muy superior al resto de componentes, por lo que se puede considerar al grafo como una componente conexa gigante, esto implica que hubo un grupo de personas muy activas sexualmente y otras menos. El promedio de los grados de los nodos es 2 quiere decir que como promedio una persona tuvo relaciones con dos mas. La distancia promedio es 3 esto implica que los lazos sexuales como promedio incluyen a 3 personas. El coeficiente de clustering promedio es 0 esto quiere decir que no hay triangulos amorosos, es decir que dos personas que estuvieron con una tercera nunca se enrollaron ellas.

b) Is this network bipartite? Justify your answer and explain its meaning for this specific network. (Caution! **is_bipartite** does not answer this question. Find the right function to answer it, or find some other way to

answer it without a predefined function.)

A simple vista se puede observar que el grafo no es bipartito, pues se observan algunas aristas entre nodos del mismo color, dando lugar a relaciones homosexuales en la red. Por esto no podemos separar los nodos en dos conjuntos al haber aristas entre nodos de un mismo tipo. Adicionalmente si se quiere chequear manualmente esto, basta con recorrer las aristas del grafo y detectar alguna conexion entre nodos con mismo color. A continuación programé una función que hace lo anterior.

```
is_bipartite2 = function(graph, attribute = 'sex') {
  edges = as_edgelist(graph, names = TRUE)

  isBipartite = TRUE  #assume bipartite

  for(i in 1:gsize(graph)) {
    firstName = edges[i,1]
    secondName = edges[i,2]

    firstAttr = vertex_attr(graph = graph, name = attribute, index = which(V(graph)$name == firstName))
    secondAttr = vertex_attr(graph = graph, name = attribute, index = which(V(graph)$name == secondName

    isBipartite = isBipartite && (firstAttr != secondAttr)
  }

  return(isBipartite)
}
```

Como resultado de evaluar el grafo en la red queda FALSE.

c) Who are the most central 3 nodes according to degree, closeness, betweenness and eigenvector centrality? Comment the differences you found, if any.

```
degree_centr = attr(sort(degree(ga, normalized=TRUE), decreasing=TRUE), "name")[1:3]
clos_centr = attr(sort(closeness(ga, normalized=TRUE), decreasing=TRUE), "name")[1:3]
betw_centr = attr(sort(betweenness(ga, normalized=TRUE), decreasing=TRUE), "name")[1:3]
eige_centr = attr(sort(eigen_centrality(ga, scale=FALSE)$vector, decreasing=TRUE), "name")[1:3]
```

Los tres nodos mas centrales por centralidad de tipo *grados* son karev, sloan, torres.

Los tres nodos mas centrales por centralidad de tipo *closeness* son torres, sloan, addison.

Los tres nodos mas centrales por centralidad de tipo *betweenness* son sloan, karev, altman.

Los tres nodos mas centrales por centralidad de tipo *eigenvector* son karev, torres, sloan.

Los nodos **karev**, **torres** y **sloan** son los que mas se repiten en los tipos de centralidades. En la de tipo *closeness* aparece **addison** ya que este nodo se encuentra mas cercano al resto que **karev**. Y en la de tipo *betweenness* aparece **altman** ya que a traves de este se conecta un subgrafo a la red mayor que el de **torres** en este caso.

d) If you wanted to test one single character for sexually transmitted infections, who would he/she be? Why?

Escogería a **torres** porque es el nodo que mayor indice de centralidad de tipo *closeness* tiene, y como resultado es el que propaga la información mas rápido a la red.

e) P. S. Bearman, J. Moody and K. Stovel (in their classic "Chains of affection: The structure of adolescent romantic and sexual networks" (*Am. J. Soc.* 110 (2004), 44-91) on romantic relations in a High School which I have mentioned several times in the course) found a prohibition against coupling with a former partner's former partner's former partner (and justified it on status implications). Without taking into account the time of the sexual contacts (which are not available in our network), what kind of structures does this rule

"forbid"? Check whether this network satisfies this prohibition or not. (Hint: Take a glance at the function **kcycle.census** in the **sna** package.)

Esta regla, según lo que entiendo, prohibe las relaciones entre un individuo y el la expareja de una de las exparejas de su pareja actual. Es decir que no se admiten ciclos de tamaño 4 en la supuesta red. Para verificar esto en la red de Anatomia de Grey usamos el siguiente código para detectar todos los ciclos en la red hasta tamaño 4.

```
net = asNetwork(ga)
net_cycles = kcycle.census(dat = net, maxlen = 4,tabulate.by.vertex = TRUE, mode = "digraph")
four_cycles_cant = net_cycles$cycle.count[3]
```

Como se puede apreciar existen 14 de tamaño 4 por lo que esta red no satisface esta prohibición.

f) Is this network consistent with the E-R model? And with the basic undirected Barabasi-Albert's model? Did you find the answers surprising?

```
set.seed(1991)
ER <- function(x,y) {
  G = sample_gnp(x, y, directed = FALSE, loops = FALSE)
  c(gorder(G),
    gsize(G),
    edge_density(G),
    components(G)$no,
    components(G)$csize[1],
    mean(degree(G)),
    mean_distance(G),
    diameter(G),
    transitivity(G, type="average"),
    transitivity(G, type="global"))
    }
X1 <- replicate(1000, ER(gorder(ga), edge_density(ga)))
er_model <- rbind(
                 c(round(apply(X1, FUN = mean, MARGIN = 1), 2)))
real_network <-  cbind(gorder(ga),
                  gsize(ga),
                  round(edge_density(ga),2),
                  components(ga)$no,
                  components(ga)$csize[1],
                  round(mean(degree(ga)),2),
                  round(mean_distance(ga),2),
                  diameter(ga),
                  transitivity(ga, type="average"),
                  transitivity(ga, type="global"))
dimnames(er_model) = list(c("E-R Model"), c("Order", "Size", "Density", "Components", "Largest Componen
dimnames(real_network) = list(c("Real Network"), c("Order", "Size", "Density", "Components", "Largest C
```

```
set.seed(1991)
BAM <- function(x, y) {
  G = sample_pa(n = x, m = y, directed = FALSE, out.pref = TRUE)
  c(gorder(G),
    gsize(G),
    edge_density(G),
    components(G)$no,
    components(G)$csize[1],
    mean(degree(G)),
```

```
    mean_distance(G),
    diameter(G),
    transitivity(G, type="average"),
    transitivity(G, type="global"))
    }
X2 <- replicate(1000, BAM(gorder(ga), 2))
bam_model <- rbind(
                c(round(apply(X2, FUN = mean, MARGIN = 1), 2)))
dimnames(bam_model) = list(c("B-A Model"), c("Order", "Size", "Density", "Components", "Largest Componer
```

```
real_network
```

	Order	Size	Density	Components	Largest Component Size	Average Degree	Average Distance	D
Real Network	44	46	0.05	4	31	2.09	3.49	

```
er_model
```

	Order	Size	Density	Components	Largest Component Size	Average Degree	Average Distance	Dia
E-R Model	44	46.07	0.05	7.24	30.38	2.09	4.13	

```
bam_model
```

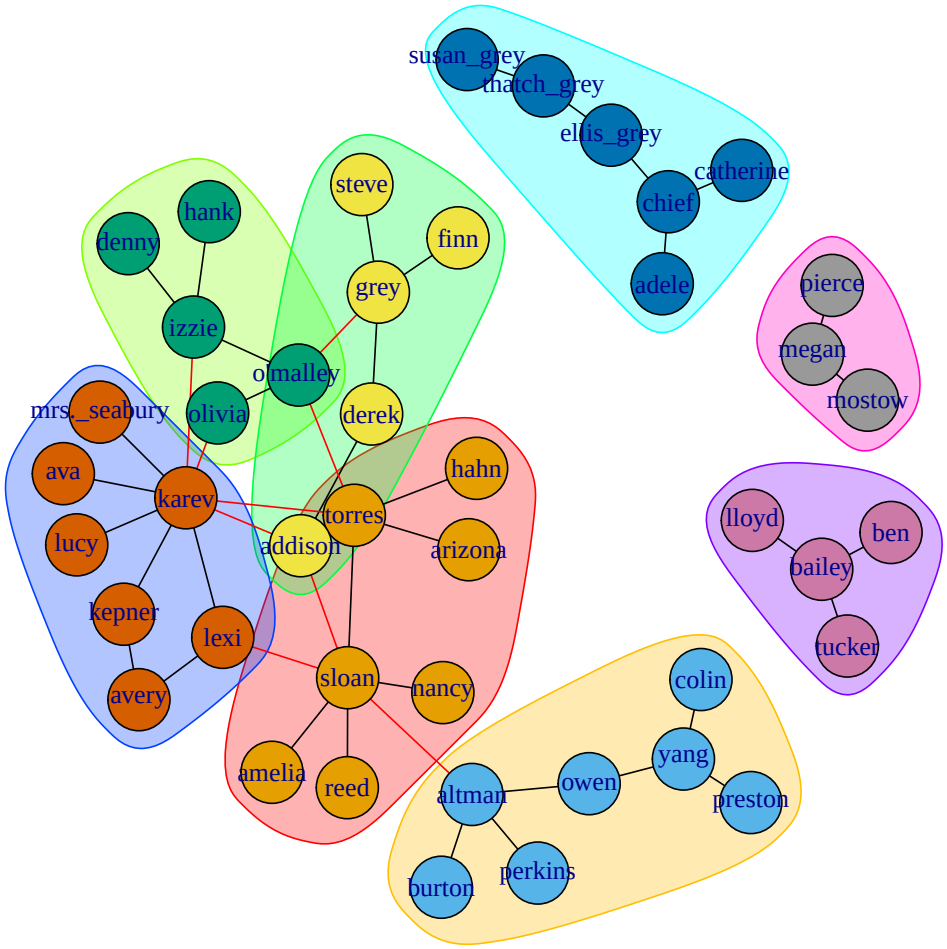
	Order	Size	Density	Components	Largest Component Size	Average Degree	Average Distance	Diar
B-A Model	44	85	0.09	1	44	3.86	2.64	

La respuesta no me sorprende tanto, pues en esta red cada nodo esta enlazado con igual probabilidad que el resto de la red. Es decir que la probabilidad de que una persona en esta red sea homosexual, bisexual o heterosexual es la misma.

Después de analizar los resultados de los dos modelos, creo que esta red es mas similar al modelo E-R y no al modelo B-A.

g) Partition the network into communities using the maximum modularity greedy algorithm (as implemented in **cluster_fast_greedy**). Plot the graph highlighting the communities. What is the modularity value of this partition? And what is the meaning of this value (in the context of this example, not the abstract definition)?

```
coords=layout_with_fr(ga)
fg=cluster_fast_greedy(ga)
plot(fg, ga, layout=coords)
```




```
ga_mod = modularity(fg)
```

En el contexto de esta red la modularidad es un indice del número de relaciones sexuales en los diferentes grupos o comunidades respecto al número de relaciones producidas por el modelo de configuración (*configuration model*). La modularidad de esta partición es 0.6583176.

h) **Extra bonus:** Are the communities found in the previous point explained by some of the nodes' attributes contained in the nodes data frame?

En algunos casos si, por ejemplo hay una comunidad conformada por solo personas de color negro, hay otras comunidades conformadas solo por personas de color blanco. En otros casos estan conformadas por personas de una misma posición, o posiciones similares.