Graph Mining

Laurent Rouvière

2021-02-07

Table des matières

Présentation

Ce tutoriel présente une introduction au **graph mining** avec **R**. On pourra trouver :

- les supports de cours associés à ce tutoriel ainsi que les données utilisées à l'adresse suivante https://lrouviere.github.io/INP-HB/ ;
- le tutoriel sans les correction à l'url https://lrouviere.github.io/TUTO_GRAPHES/
- le tutoriel avec les corrigés (à certains moment) à l'url https://lrouviere.github.io/TUTO_GRAPHES/correction/.

Il est recommandé d'utiliser **mozilla firefox** pour lire le tutoriel.

Les thèmes suivants sont abordés :

- Manipulation de graphes avec **igraph** : visualisation statique et dynamique
- **Modèles** basiques pour des graphes, notamment Erdos-Renyi et SBM
- **Détection de communautés**, maximisation de la modularité
- **Clustering spectral**, pour des graphes mais aussi des "données standards".

1 Manipulation de graphes avec igraph

Le but de ce tutoriel est de se familiariser avec les principales fonctions du package **igraph**. On trouvera un descriptif de ce package à l'adresse http://igraph.org/r/. On pourra également consulter le tutoriel (très complet) suivant : http://kateto.net/networks-r-igraph.

Nous commençons par charger le package

```
library(igraph)
```

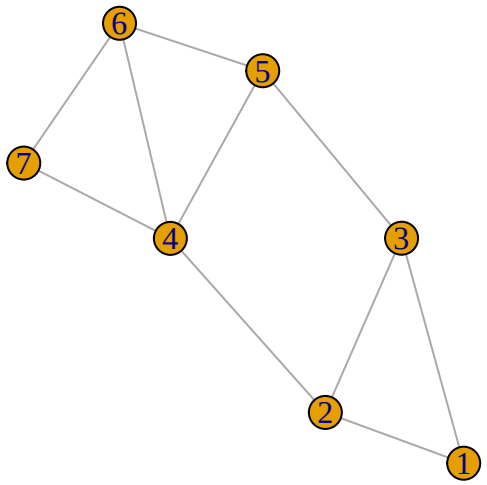
1.1 Construction de graphes avec igraph

Comme pour des données classiques, il est possible de construire des graphes directement dans **R** ou de les importer à partir de fichiers externes.

1.1.1 Quelques fonctions R pour construire des graphes

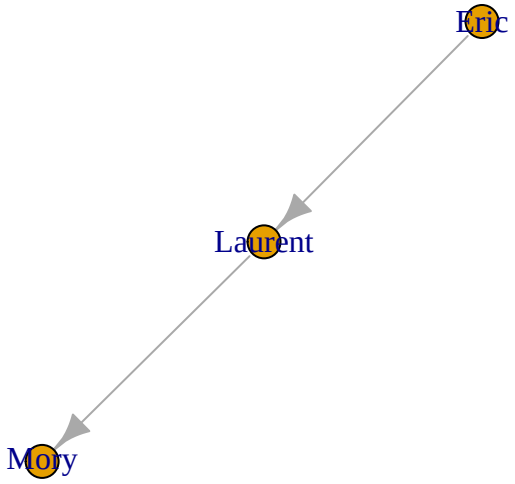
L'approcher la plus naturelle est de définir un graphe à partir d'une liste d'arêtes :

```
g1 <- graph(edges=c(1,2,1,3,2,3,3,5,2,4,4,5,5,6,4,6,4,7,6,7),n=7,directed=F)
plot(g1)
```



Si la liste d'arêtes est donnée sous forme de noms, il n'est pas nécessaire d'indiquer le nombre de nœuds.

```
g2 <- graph(edges=c("Eric","Laurent","Laurent","Mory"))
plot(g2)
```

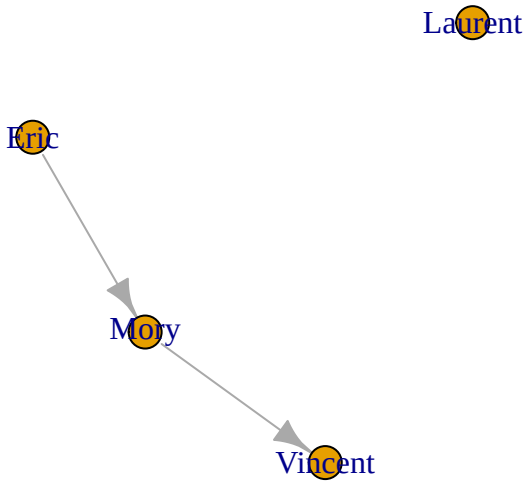


Différentes options sont proposées dans la fonction **graph** :

- `n` : le nombre maximal de nœuds
- `isolates` : ajout de nœuds isolés
- `directed` : graphes dirigés ou non
- ...

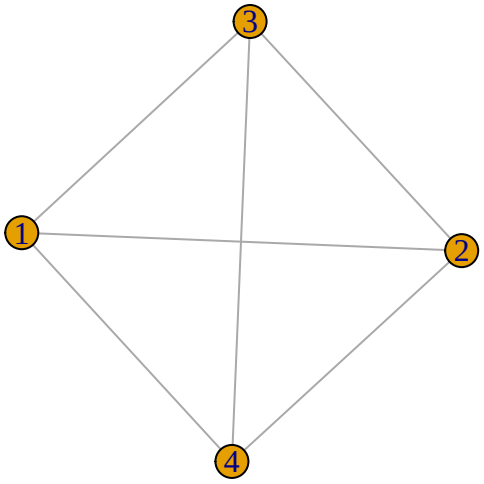
Par exemple

```
g3 <- graph(edges=c("Eric","Mory","Mory","Vincent"),isolates="Laurent")
plot(g3)
```

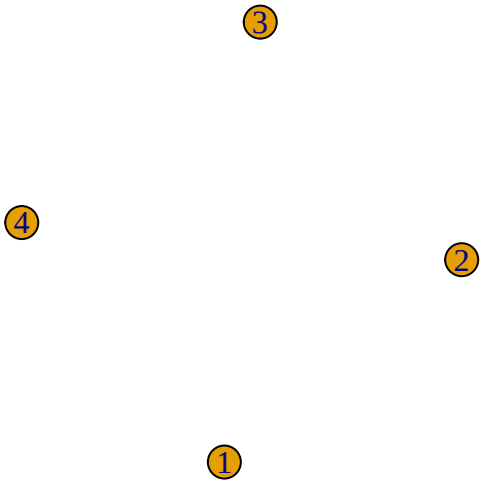


Il existe également des fonctions spécifiques qui peuvent aider à la construction de graphes, par exemple **make_full_graph** :

```
plot(make_full_graph(4))
```

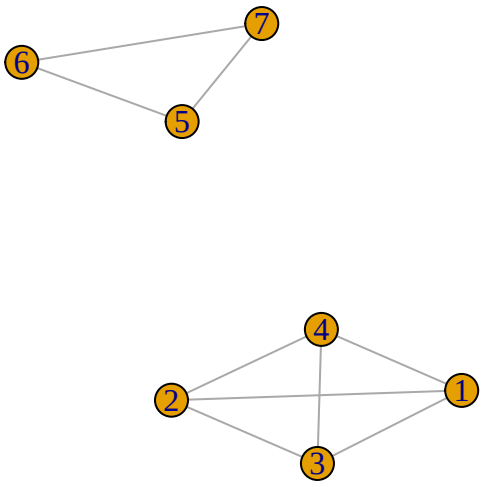



```
plot(make_empty_graph(4))
```



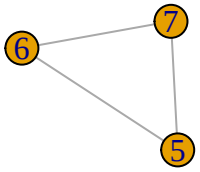
Il est également possible “d’additionner” des graphes

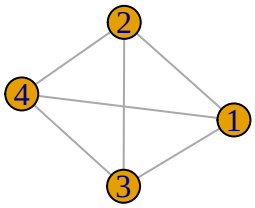
```
plot(make_full_graph(4)+make_full_graph(3))
```



L'opérateur **pipe** permet une lecture du code plus lisible

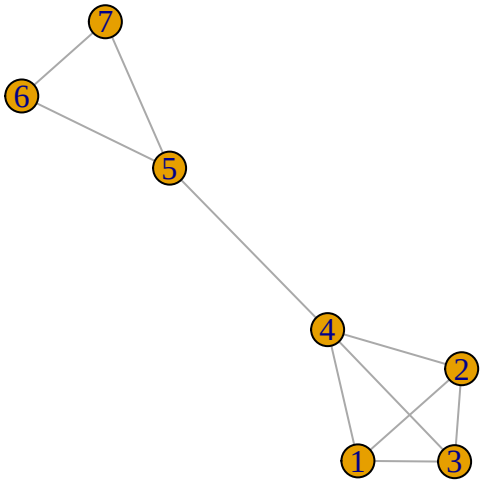
```
(make_full_graph(4)+make_full_graph(3)) %>% plot()
```





Il est également facile d'ajouter des arêtes avec **add_edges** :

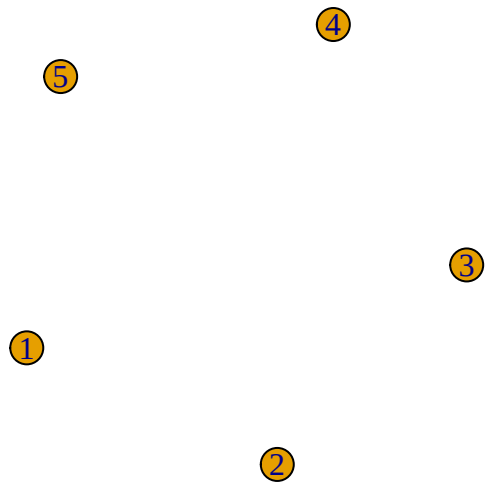
```
(make_full_graph(4)+make_full_graph(3)) %>% add_edges(c(4,5)) %>% plot()
```



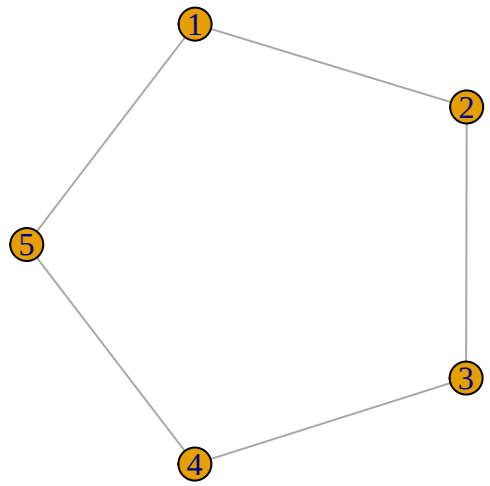
Exercice 1.1 (Quelques graphes spécifiques)**.**

Tester et expliquer les fonctions **make_empty_graph, make_ring** et **make_star**.

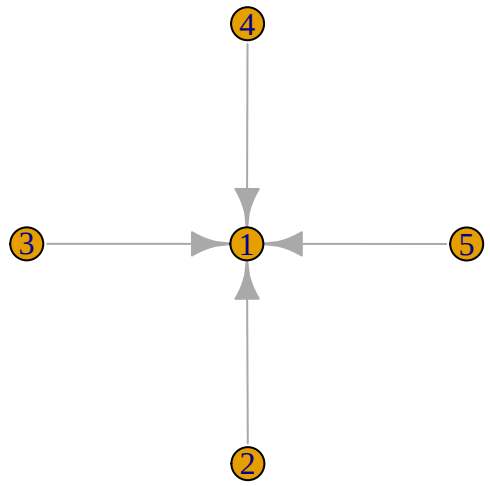
```
make_empty_graph(5) %>% plot()
```




```
make_ring(5) %>% plot()
```




```
make_star(5) %>% plot()
```



On remarque que :

— ***make_empty_graph*** *produit un graphe sans arête ;*
— ***make_ring*** *donne un graphe en cercle ;*
— ***make_star*** *donne un graphe en étoile : tous les nœuds sont connectés à un même nœud.*

1.1.2 Construction à partir d'un fichier externe

Le plus souvent, on aura à récupérer des données récoltées dans des fichiers `txt` ou `csv` pour construire le graphe.

Exercice 1.2 (Importation)**.**

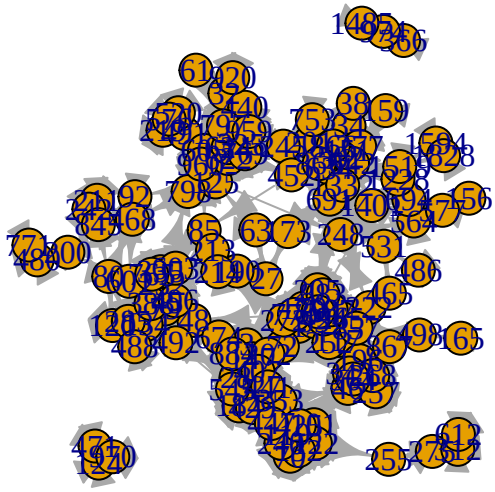
On considère le jeu de données **Friendship-network_data_2013.csv** qui se trouve sur le site http://www.sociopatterns.org/datasets/high-school-contact-and-friendship-networks/. Ces données concernent des relations entre étudiants.

1. Importer les données à l'aide de **read.table**.

```
friends <- read.table(file='data/Friendship-network_data_2013.csv')
head(friends)
  V1  V2
1  1  55
2  1 205
3  1 272
4  1 494
5  1 779
6  1 894
```

2. Ce fichier contient 2 colonnes et chaque colonne contient une arête. Visualiser le graphe. On pourra utiliser **graph_from_data_frame**.

```
amis <- graph_from_data_frame(friends,directed=T)
amis
IGRAPH 9a34439 DN-- 134 668 --
+ attr: name (v/c)
+ edges from 9a34439 (vertex names):
 [1] 1 ->55  1 ->205 1 ->272 1 ->494 1 ->779 1 ->894 3 ->1
 [8] 3 ->28  3 ->147 3 ->272 3 ->407 3 ->674 3 ->884 27->63
[15] 27->173 28->202 28->327 28->353 28->407 28->429 28->441
[22] 28->492 28->545 32->440 32->624 32->797 32->920 34->151
[29] 34->277 34->502 34->866 45->48  45->79  45->335 45->496
[36] 45->601 45->674 45->765 46->117 46->196 46->257 46->268
[43] 48->45  48->79  48->496 55->1   55->170 55->205 55->252
[50] 55->272 55->779 55->883 55->894 61->797 63->27  63->125
+ ... omitted several edges
plot(amis)
```



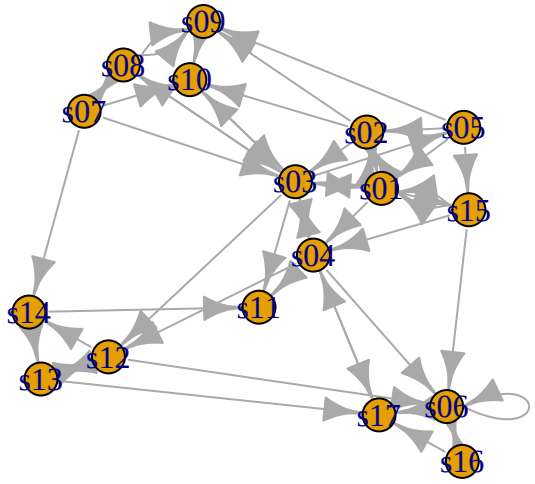
3. On considère un graphe permettant de visualiser des connexions entre médias :
   - les nœuds sont définis dans le fichier `Dataset1-Media-Example-NODES.csv`
   - les arêtes dans le fichier `Dataset1-Media-Example-EDGES.csv`.

   Importer ces fichiers.

```
nodes <- read.csv("data/Dataset1-Media-Example-NODES.csv", header=T)
links <- read.csv("data/Dataset1-Media-Example-EDGES.csv", header=T)
head(nodes)
   id               media media.type type.label
1 s01            NY Times          1  Newspaper
2 s02     Washington Post          1  Newspaper
3 s03 Wall Street Journal          1  Newspaper
4 s04           USA Today          1  Newspaper
5 s05            LA Times          1  Newspaper
6 s06       New York Post          1  Newspaper
  audience.size
1            20
2            25
3            30
4            32
5            20
6            50
head(links)
  from  to weight      type
1  s01 s02     10 hyperlink
2  s01 s02     12 hyperlink
3  s01 s03     22 hyperlink
4  s01 s04     21 hyperlink
5  s04 s11     22   mention
6  s05 s15     21   mention
```

4. Construire le graphe `igraph` associé à ces deux fichiers à l'aide de **graph_from_data_frame**.

```
net <- graph_from_data_frame(d=links, vertices=nodes, directed=T)
plot(net)
```

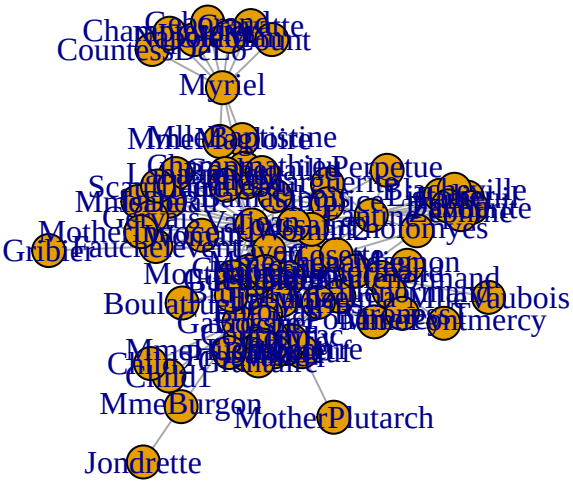


5. La fonction **read_graph** peut s'adapter à de nombreux formats de graphe :

```
read_graph(file, format = c("edgelist", "pajek", "ncol", "lgl", "graphml", "dimacs", "graphdb", "gm
```

On considère par exemple le fichier **lesmis.gml** disponible ici. Les nœuds correspondent aux personnages du roman et une arête est présente si deux personnages apparaissent dans le même chapitre. Le poids de l'arête est déterminé par le nombre de chapitres où les deux personnages sont présents. Importer le graphe à l'aide de **read_graph** et visualiser le.

```
miserab <- read.graph(file='data/lesmis.gml',format="gml")
plot(miserab)
```



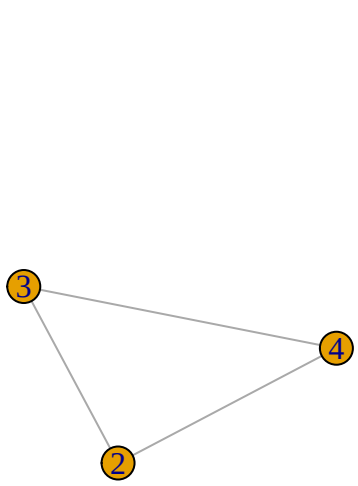
1.1.3 Matrice d'adjacence

Enfin un graphe peut également s'identifier avec une **matrice d'adjacence** :

```
A <- matrix(c(0,0,0,0,0,0,1,1,0,1,0,1,0,1,1,0), 4, 4)
A
     [,1] [,2] [,3] [,4]
[1,]    0    0    0    0
[2,]    0    0    1    1
[3,]    0    1    0    1
[4,]    0    1    1    0
```

On pourra utiliser dans ce cas la fonction **graph_from_adjacency_matrix** pour convertir la matrice en un objet `igraph` :

```
G <- graph_from_adjacency_matrix(A, mode='undirected')
plot(G)
```



On peut bien entendu faire l'opération inverse et calculer la **matrice d'ajacence** d'un graphe :

```
as_adj(G)
4 x 4 sparse Matrix of class "dgCMatrix"

[1,] . . . .
[2,] . . 1 1
[3,] . 1 . 1
[4,] . 1 1 .
```

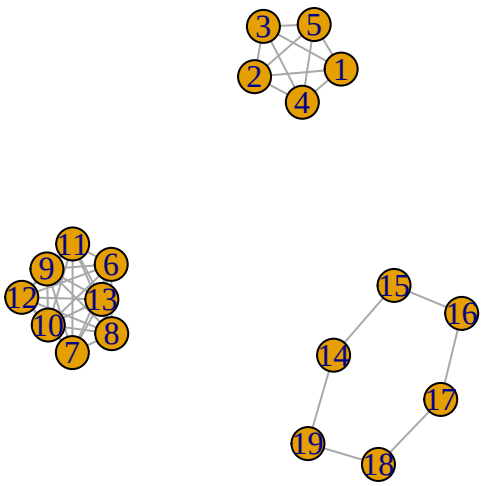
1.2 Visualisation d'un graphe

Un des intérêts principaux du graph mining et de visualiser les connexions entre les nœuds à l'aide d'un graphe. Se pose bien entendu la question (difficile) de la position des nœuds dans le plan pour obtenir la visualisation la plus pertinente du graphe. On peut ensuite s'interroger sur des outils classiques qui vont permettre de colorier les nœuds, d'utiliser différents symboles pour les arêtes, etc...

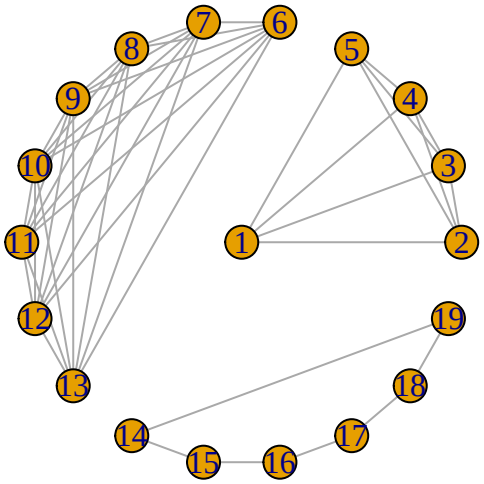
1.2.1 Network layouts : algorithmes usuels de visualisation

Un graphe peut être visualisé à l'aide de plusieurs algorithmes. On pourra trouver un descriptif ici. On se contentera de donner différents layouts pour l'exemple suivant :

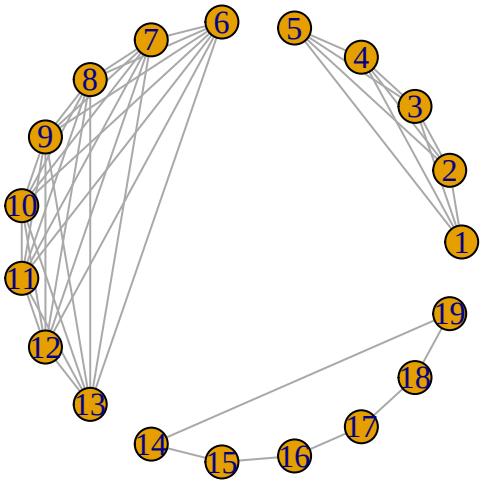
```
G <- make_full_graph(5)+make_full_graph(8)+make_ring(6)
plot(G)
```



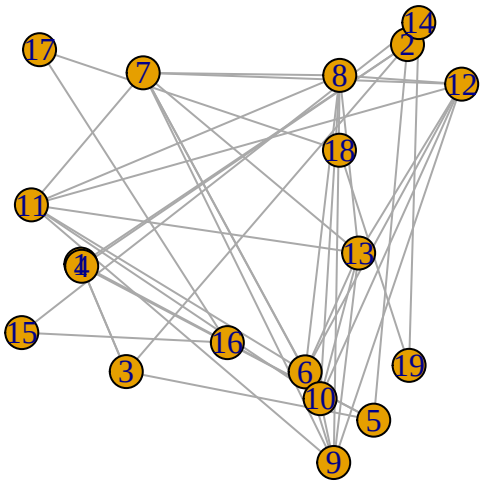

```
plot(G,layout=layout_as_star(G))
```




```
plot(G,layout=layout.circle(G))
```

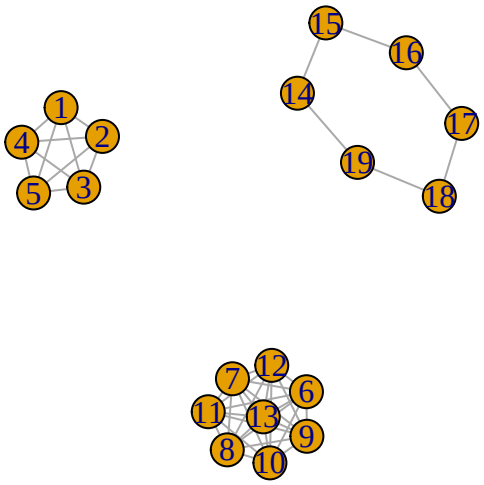


```
plot(G,layout=layout_randomly(G))
```

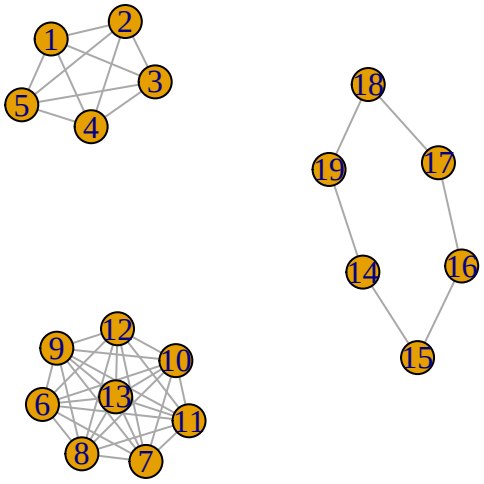


Deux algorithmes sont connus pour avoir des visualisations jugées “esthétiques”. L’idée, très rapidement, est d’essayer d’obtenir la position des nœuds et des arêtes de façon uniforme dans le plan. Pour plus d’informations sur ce sujet difficile on pourra consulter cet article.

```
plot(G,layout=layout_with_fr(G))
```




```
plot(G,layout=layout_with_kk(G))
```



Enfin la fonction **tktplot**

```
tkplot(G)
```

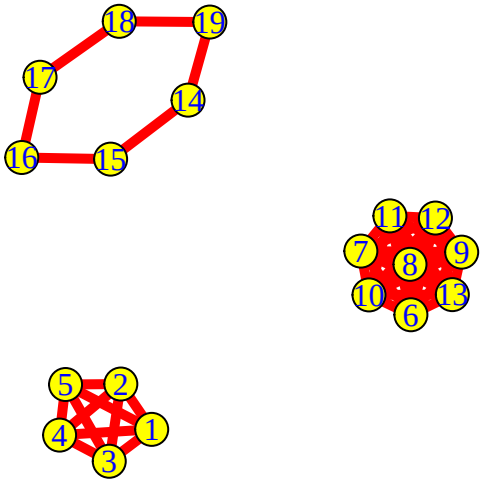
1.2.2 Personnalisation du graphe

Il est bien entendu possible de modifier les couleurs, tailles... des nœuds et arêtes. Deux stratégies sont possibles avec **igraph** :

- utiliser les options de **plot.igraph** : vertex.color, vertex.shape, vertex.size, vertex.label... et edge.color, edge.label, edge.width...
- travailler sur les nœuds et arêtes séparément à l'aide des fonctions **V()** et **E()**.

La fonction **plot.igraph** :

```
plot(G,
     vertex.color="yellow",vertex.size=15,vertex.label.color="blue",
     edge.color="red",edge.width=5)
```



On peut également modifier les paramètre des nœuds

```
G1 <- G
V(G1)$color <- "red"
V(G1)$size <- 15
```

et des arêtes

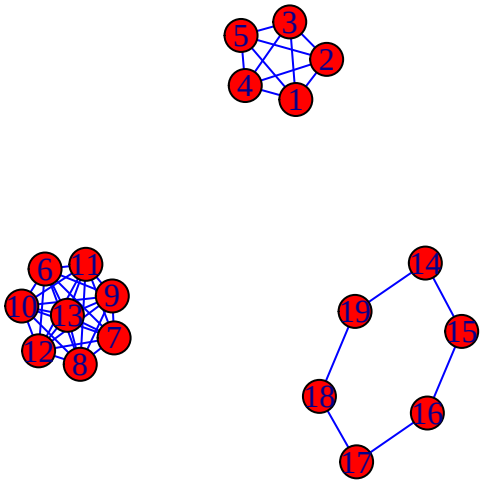
```
E(G1)$color <- "blue"
E(G1)$size <- 3
```

On a ainsi

```
vertex_attr(G1)
$color
 [1] "red" "red" "red" "red" "red" "red" "red" "red" "red"
[10] "red" "red" "red" "red" "red" "red" "red" "red" "red"
[19] "red"

$size
 [1] 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15
[19] 15
edge_attr(G1)
$color
 [1] "blue" "blue" "blue" "blue" "blue" "blue" "blue" "blue"
 [9] "blue" "blue" "blue" "blue" "blue" "blue" "blue" "blue"
[17] "blue" "blue" "blue" "blue" "blue" "blue" "blue" "blue"
[25] "blue" "blue" "blue" "blue" "blue" "blue" "blue" "blue"
[33] "blue" "blue" "blue" "blue" "blue" "blue" "blue" "blue"
[41] "blue" "blue" "blue" "blue"

$size
 [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[29] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

et on peut visualiser le graphe (sans option dans **plot.igraph**) :

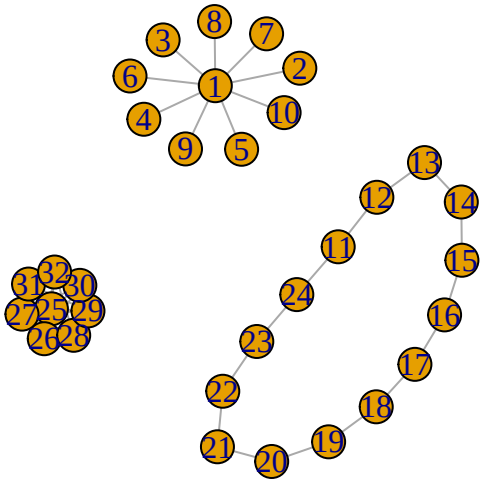
```
plot(G1)
```



Exercice 1.3 (Gérer les couleurs avec igraph)**.**

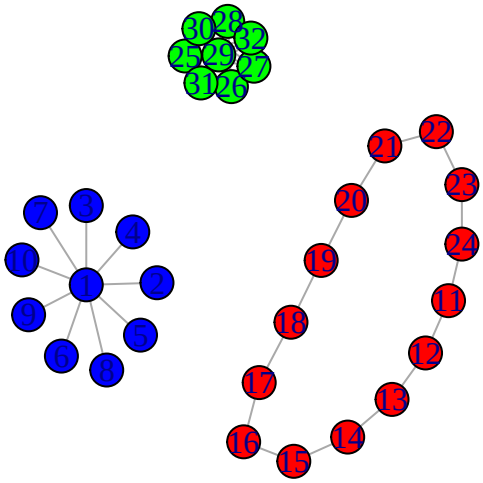
1. Construire un graphe avec 3 composantes composantes connexes de taille 10, 14 et 8.

```
G <- make_star(10,mode="undirected")+make_ring(14)+make_full_graph(8)
plot(G)
```



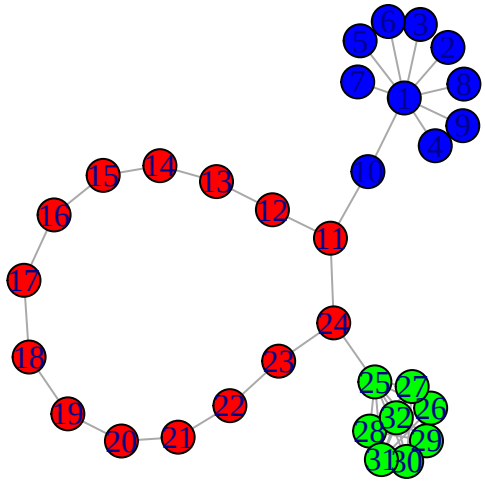
2. Colorier les nœuds de chaque composante d'une couleur différente.

```
V(G)$color <- c(rep("blue",10),rep("red",14),rep("green",8))
plot(G)
```



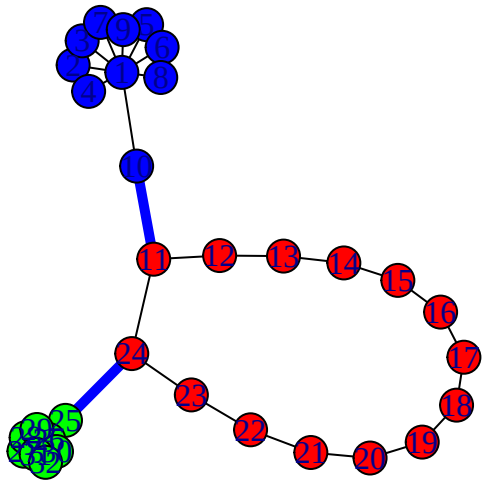
3. Relier les composantes en ajoutant 2 arêtes (et pas plus).

```
G1 <- add_edges(G,c(10,11,24,25))
plot(G1,layout=layout_with_kk(G1))
```



4. Utiliser une couleur et une taille différente pour les deux arêtes crées.

```
nb.arr <- length(E(G1))
E(G1)$color <- c(rep("black",nb.arr-2),"blue","blue")
E(G1)$width <- c(rep(1,nb.arr-2),5,5)
plot(G1)
```



Exercice 1.4 (Customiser un graphe)**.**

On considère le graphe **net** sur les médias défini dans l'exercice 1.2. Représenter le graphe en ajoutant :

— le nom des nœuds (**media**)
— une couleur différente en fonction du type de média
— une taille de nœud différente en fonction de l'audience
— une taille d'arête différente en fonction du poids (`weight`)
— une couleur d'arête différente en fonction du type (`type`).

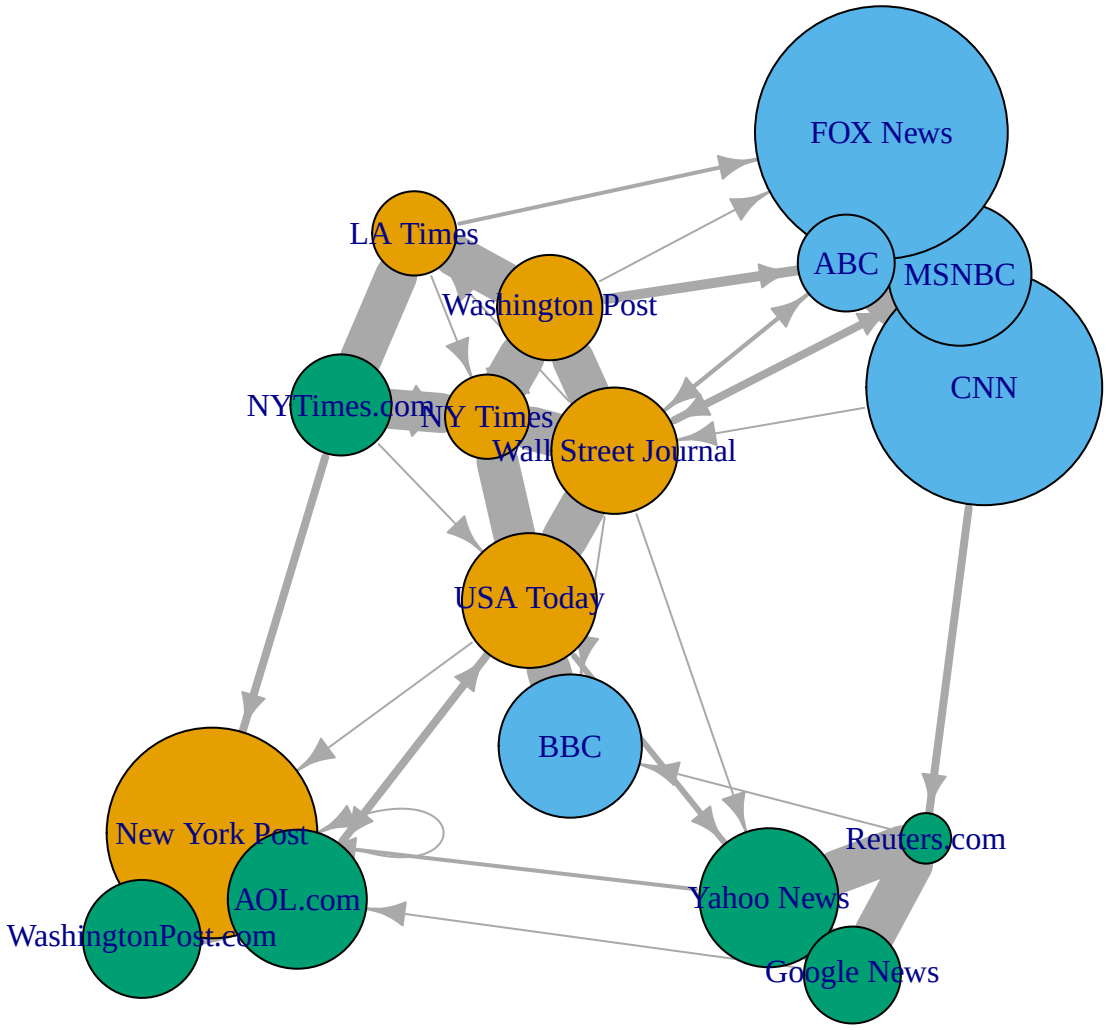
On regarde tout d'abord les attributs des noeuds et arêtes du graphe :

```
vertex_attr(net)
$name
 [1] "s01" "s02" "s03" "s04" "s05" "s06" "s07" "s08" "s09"
[10] "s10" "s11" "s12" "s13" "s14" "s15" "s16" "s17"
```

```
$media
 [1] "NY Times"             "Washington Post"
 [3] "Wall Street Journal" "USA Today"
 [5] "LA Times"             "New York Post"
 [7] "CNN"                  "MSNBC"
 [9] "FOX News"             "ABC"
[11] "BBC"                  "Yahoo News"
[13] "Google News"          "Reuters.com"
[15] "NYTimes.com"          "WashingtonPost.com"
[17] "AOL.com"

$media.type
 [1] 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 3

$type.label
 [1] "Newspaper" "Newspaper" "Newspaper" "Newspaper"
 [5] "Newspaper" "Newspaper" "TV"         "TV"
 [9] "TV"        "TV"        "TV"         "Online"
[13] "Online"    "Online"    "Online"     "Online"
[17] "Online"

$audience.size
 [1] 20 25 30 32 20 50 56 34 60 23 34 33 23 12 24 28 33
edge_attr(net)
$weight
 [1] 10 12 22 21 22 21 21 11 12 22 23 20 11 11 21 23 21 21
[19] 21 22 21 21 21 23 21 22 22 21 21  2  5  1  1  2  2  3
[37]  1  1  1  1  2  2  4  2  4  4  4  1  1  1  1  1

$type
 [1] "hyperlink" "hyperlink" "hyperlink" "hyperlink"
 [5] "mention"   "mention"   "mention"   "mention"
 [9] "mention"   "hyperlink" "hyperlink" "mention"
[13] "hyperlink" "hyperlink" "mention"   "hyperlink"
[17] "hyperlink" "mention"   "mention"   "mention"
[21] "hyperlink" "hyperlink" "hyperlink" "hyperlink"
[25] "hyperlink" "hyperlink" "mention"   "mention"
[29] "hyperlink" "hyperlink" "hyperlink" "hyperlink"
[33] "mention"   "hyperlink" "mention"   "hyperlink"
[37] "mention"   "hyperlink" "mention"   "hyperlink"
[41] "mention"   "mention"   "hyperlink" "hyperlink"
[45] "hyperlink" "mention"   "hyperlink" "hyperlink"
[49] "mention"   "hyperlink" "hyperlink" "mention"
```

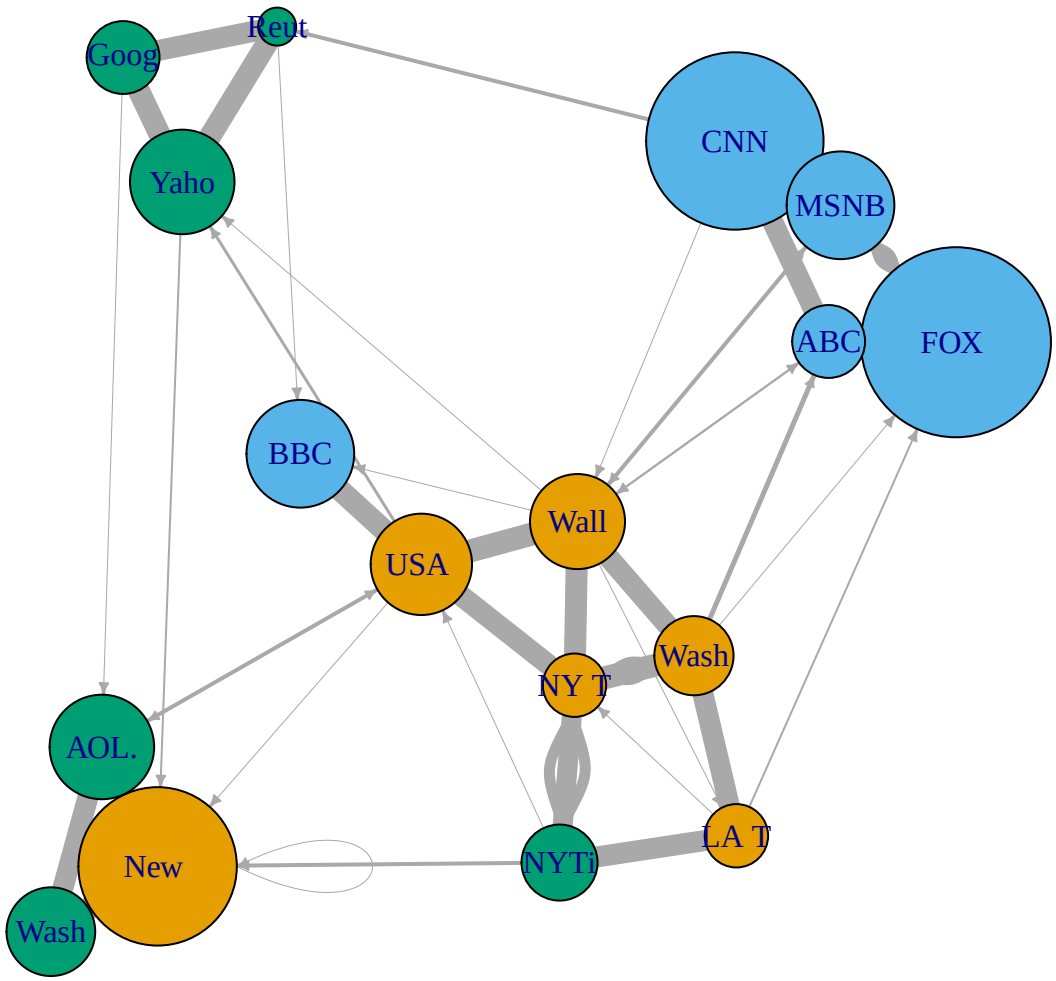
On effectue une première représentation avec :

```
plot(net,vertex.label=V(net)$media, vertex.color=V(net)$media.type,
     vertex.size=V(net)$audience.size, edge.width=E(net)$weight)
```



On peut améliorer l'esthétique du graphe en redéfinissant les épaisseurs des arêtes ou les tailles des flèches ou encore en mettant des abréviations pour les noms de média...

```
vertex_attr(net)$abbr.media <- vertex_attr(net)$media %>% str_sub(1,4)
plot(net,  vertex.label=V(net)$abbr.media, vertex.color=V(net)$media.type,
     vertex.size=0.7*V(net)$audience.size,edge.width=0.5*E(net)$weight,
     edge.arrow.size=0.3,edge.arrow.width=1.2)
```



1.3 Statistiques descriptives sur les graphes

Comme pour tout les types de données, il est souvent crucial de calculer des indicateurs descriptifs sur les graphes. Nous présentons les indicateurs standards tels que le diamètre, la densité, les degrés de centralité et d'intermédiarité...

Exercice 1.5 (Quelques descripteurs)**.**

On considère toujours le graphe des exercices précédents sur les média (`net`).

1. Calculer les nombre de nœuds, d'arêtes, le diamètre et la densité du graphe.

```
vcount(net)
[1] 17
ecount(net)
[1] 52
diameter(net)
[1] 75
edge_density(net)
[1] 0.1911765
```

2. Combien y a t-il de triangles dans le graphes ? On pourra calculer ce nombre de plusieurs façons.
   La fonction

```
C_tr <- count_triangles(net)
C_tr
 [1] 6 6 9 7 5 4 2 1 2 3 1 3 1 1 3 1 2
```

renvoie, pour chaque noeud, le nombre de triangles dont il fait partie. On déduit ainsi que le nombre de triangles vaut

```
sum(C_tr)/3
[1] 19
```

On peut aussi l'obtenir avec

```
length(triangles(net))/3
[1] 19
```

ou encore en cherchant les cliques de taille 3 :

```
length(cliques(net,min=3,max=3))
[1] 19
```

3. Calculer la transitivité.

```
transitivity(net,type="global")
[1] 0.372549
```

4. Quels sont les nœuds connectés avec le nœud 3 ? On pourra utiliser **neighbors**.

```
neighbors(net,3)
+ 7/17 vertices, named, from d41d746:
[1] s01 s04 s05 s08 s10 s11 s12
```

5. Étudier les composantes connexes du graphe.

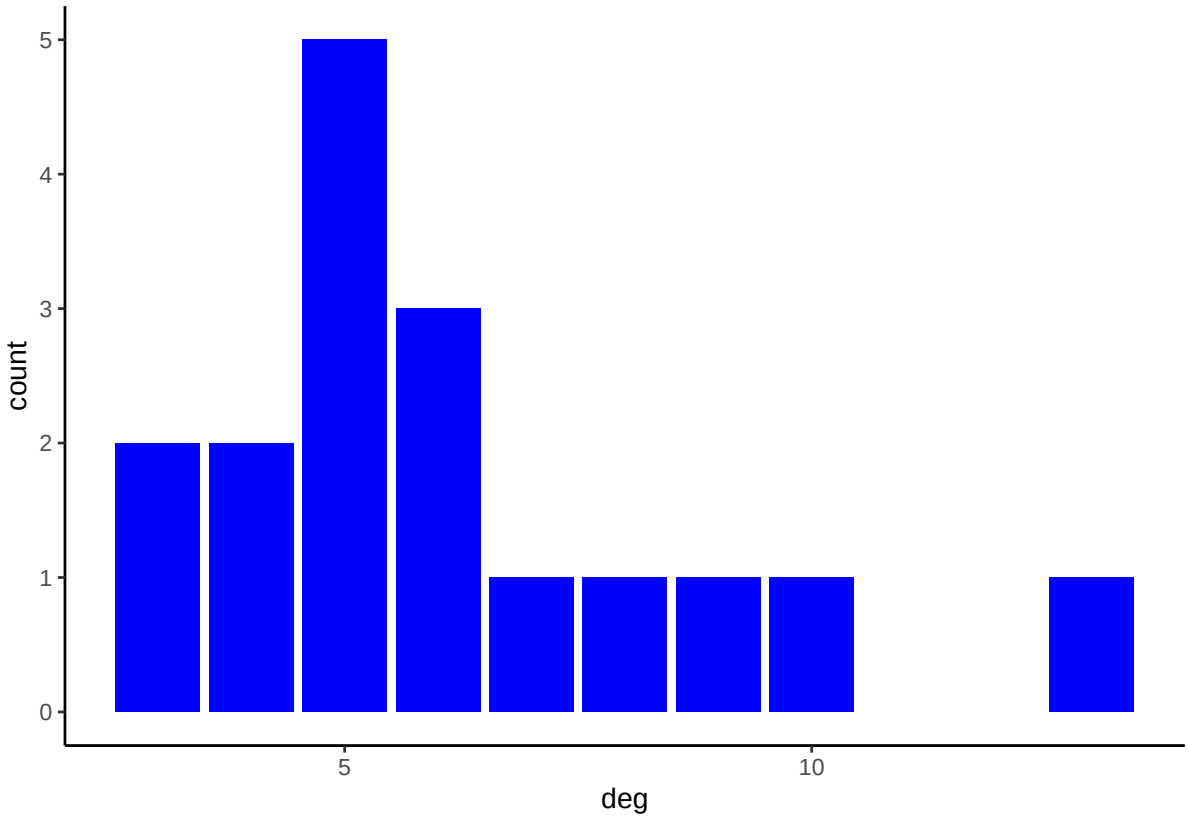
```
components(net)
$membership
s01 s02 s03 s04 s05 s06 s07 s08 s09 s10 s11 s12 s13 s14 s15
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
s16 s17
  1   1

$csize
[1] 17

$no
[1] 1
```

C'est facile : il y en a une seule de taille 17.

6. Calculer les degrés des nœuds et représenter les avec un barplot. On pourra utiliser **degree** puis **degree_distribution**.

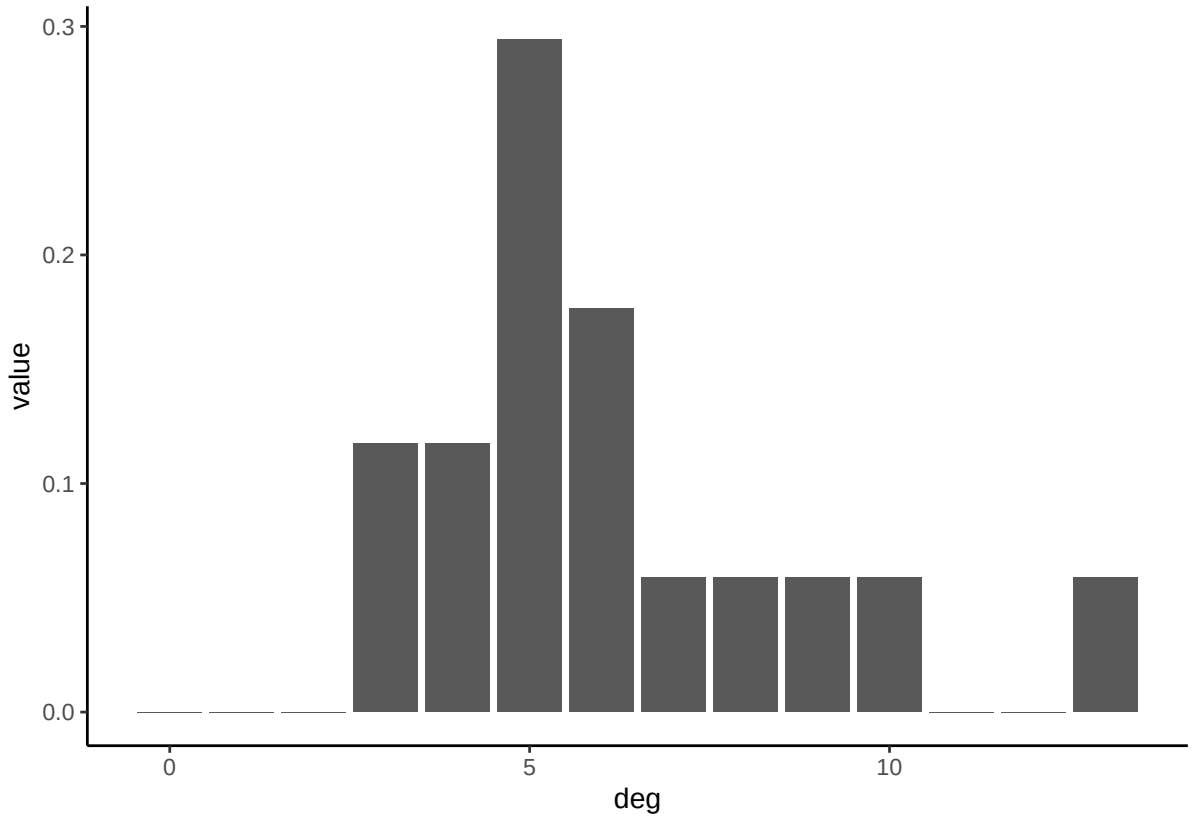
```
deg <- degree(net)
deg
s01 s02 s03 s04 s05 s06 s07 s08 s09 s10 s11 s12 s13 s14 s15
 10   7  13   9   5   8   5   6   5   5   3   6   4   4   6
s16 s17
  3   5
```

```
df <- data.frame(deg=deg)
library(tidyverse)
ggplot(df)+aes(x=deg)+geom_bar(fill="blue")+theme_classic()
```



Pour le barplot, il est préférable d'utiliser la distribution des degrés :

```
df1 <- data.frame(deg=0:13,value=degree_distribution(net))
ggplot(df1)+aes(x=deg,y=value)+geom_bar(stat="identity")
```



7. Calculer les degrés de **proximité** et d'**intermédiarité** et ordonner les observations en fonction de ces degrés.

```
prox <- closeness(net)
inter <- betweenness(net)
```

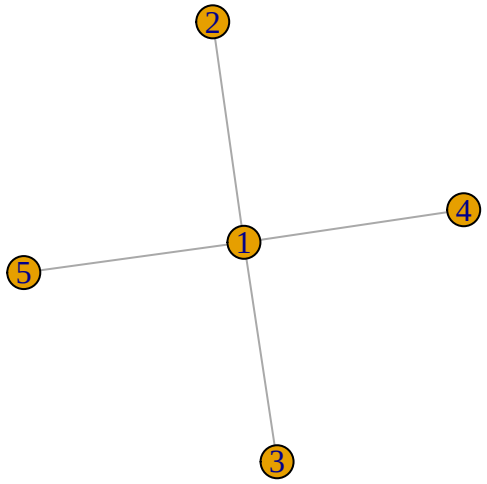
```
df.deg <- data.frame(prox,ord.prox=1:vcount(net)) %>%
  arrange(desc(prox))
df.int <- data.frame(inter,ord.inter=1:vcount(net)) %>%
  arrange(desc(inter))
bind_cols(df.deg,df.int)
           prox ord.prox      inter ord.inter
s07 0.006134969        7 145.833333         3
s03 0.005208333        3  91.333333         4
s08 0.004672897        8  66.000000        17
s10 0.004504505       10  49.000000         5
s11 0.003676471       11  48.000000        10
s02 0.003584229        2  45.500000        12
s15 0.003184713       15  34.000000         1
s05 0.002754821        5  23.666667         2
s01 0.002666667        1  21.500000         6
s04 0.002645503        4  21.500000        13
s13 0.002341920       13  14.000000         8
s17 0.002272727       17   2.333333        15
s09 0.001808318        9   1.000000        14
s12 0.001508296       12   0.500000         7
s14 0.001457726       14   0.000000         9
```

```
s06 0.001342282        6   0.000000        11
s16 0.001338688       16   0.000000        16
```

Exercice 1.6 (Centralité et intermédiarité)**.**

On considère le graphe suivant.

```
G <- make_star(5,mode="undirected")
plot(G)
```



Calculer en utilisant la définition les degrés de centralité et d'intermédiarité des nœuds de G. Retrouver ces valeurs à l'aide de fonctions **R**.

Pour la ***centralité*** *:*

- *le noeud 1 est à une distance 1 des quatre autres, son degré est donc 1/4.*
- *2 est à une distance 1 de 1, et à distance 2 des autres. Son degré est donc 1/7*
- *idem pour 3 et 4.*

*Pour l'****intermédiarité*** *:*

- *Tous les plus courts chemins passent par 1 ! Le degrés est donc égal au nombre de couples de noeuds, c'est-à-dire 6 !*
- *Pour les autres points, on remarque qu'aucun plus court chemin ne passe par eux, leur degré est donc nul.*

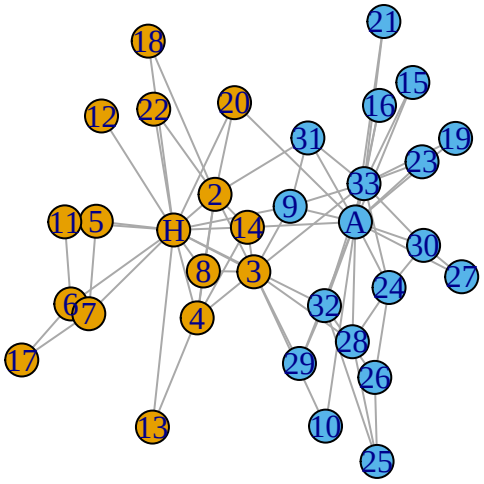
On peut bien entendu retrouver ces résultats :

```
closeness(G)
[1] 0.2500000 0.1428571 0.1428571 0.1428571 0.1428571
betweenness(G)
[1] 6 0 0 0 0
```

Exercice 1.7 (Comparaison de nœuds)**.**

On considère le graphe **karate** que l'on peut récupérer dans le package **igraphdata** :

```
library(igraphdata)
data(karate)
plot(karate)
```



En étudiant les différents critères d'importance des nœuds, identifier les nœuds importants et interpréter.

```
deg <- degree(karate)
prox <- closeness(karate)
inter <- betweenness(karate)
df.deg <- tibble(deg=deg,prox=prox,inter=inter)
deg.rank <- tibble(deg=order(deg,decreasing = TRUE),
                   prox=order(prox,decreasing = TRUE),
                   inter=order(inter,decreasing = TRUE))
deg.rank
# A tibble: 34 x 3
     deg  prox inter
   <int> <int> <int>
 1    34     1     1
 2     1    34    34
 3    33    20    20
 4     3    13    32
 5     2    21    33
 6     4    29     3
 7    32    32    25
 8     9    33     2
 9    14     9    18
10    24     3     6
# ... with 24 more rows
```

Les noeuds 1 et 34 se trouvent toujours aux deux premières places, ce sont les deux présidents des clubs : les liens avec les membres de leurs clubs sont forts. L'individu 20 est également très bien placé, on remarque qu'il est lié aux deux présidents, c'est donc un intermédiaire très important.

```
neighbors(karate,20)
+ 3/34 vertices, named, from 4b458a1:
[1] Mr Hi   Actor 2 John A
```

C'est également le cas du 32 :

```
neighbors(karate,32)
+ 6/34 vertices, named, from 4b458a1:
[1] Mr Hi    Actor 25 Actor 26 Actor 29 Actor 33 John A
```

On remarquera enfin que le 33 possède un grand nombre de liens, et donc un degrés élevé :

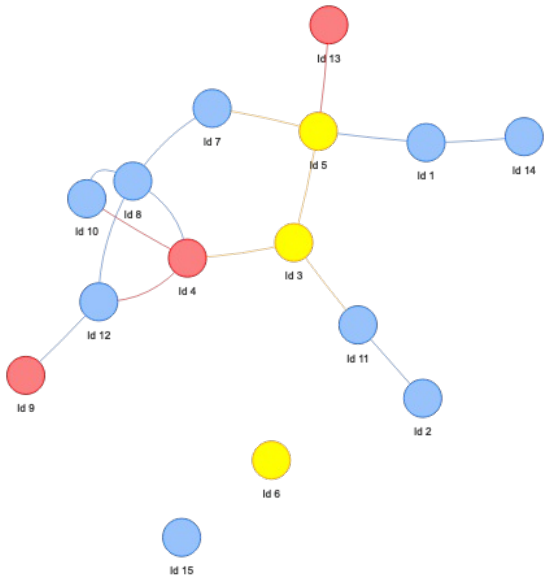
```
neighbors(karate,33)
+ 12/34 vertices, named, from 4b458a1:
 [1] Actor 3  Actor 9  Actor 15 Actor 16 Actor 19 Actor 21
 [7] Actor 23 Actor 24 Actor 30 Actor 31 Actor 32 John A
```

1.4 Autres packages pour visualiser les graphes

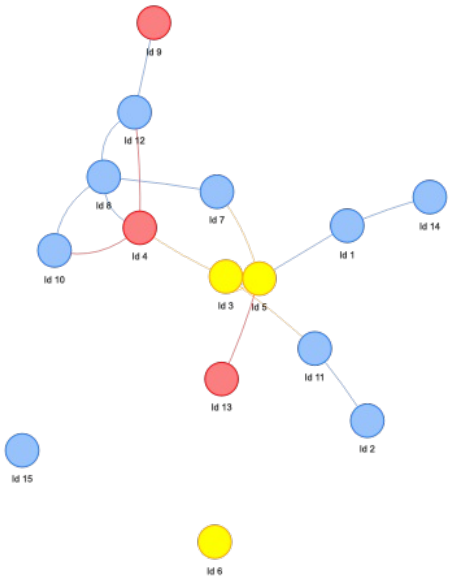
1.4.1 Graphes dynamiques avec visNetwork

Nous avons vu que le package igraph propose une visualisation statique d'un réseau. Pour donner un caractère dynamique à ce type de représentation, on pourra utiliser le package visNetwork. Une représentation standard **visNetwork** s'effectue en spécifiant les nœuds et connexions d'un graphe. Voici quelques exemples d'utilisation.

```
set.seed(1234)
nodes <- data.frame(id = 1:15, label = paste("Id", 1:15),
                    group=sample(LETTERS[1:3], 15, replace = TRUE))
edges <- data.frame(from = trunc(runif(15)*(15-1))+1,to = trunc(runif(15)*(15-1))+1)
library(visNetwork)
visNetwork(nodes,edges)
```

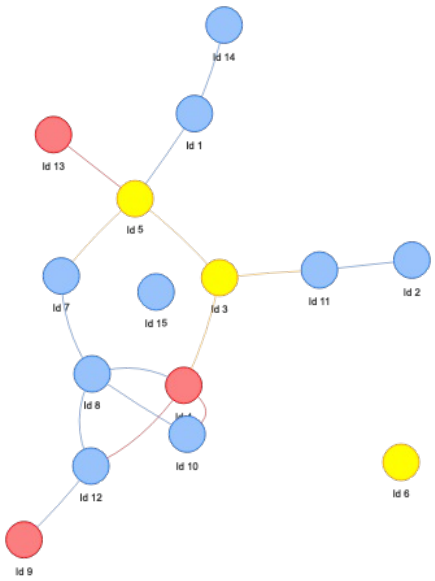



```
visNetwork(nodes, edges) %>% visOptions(highlightNearest = TRUE)
```

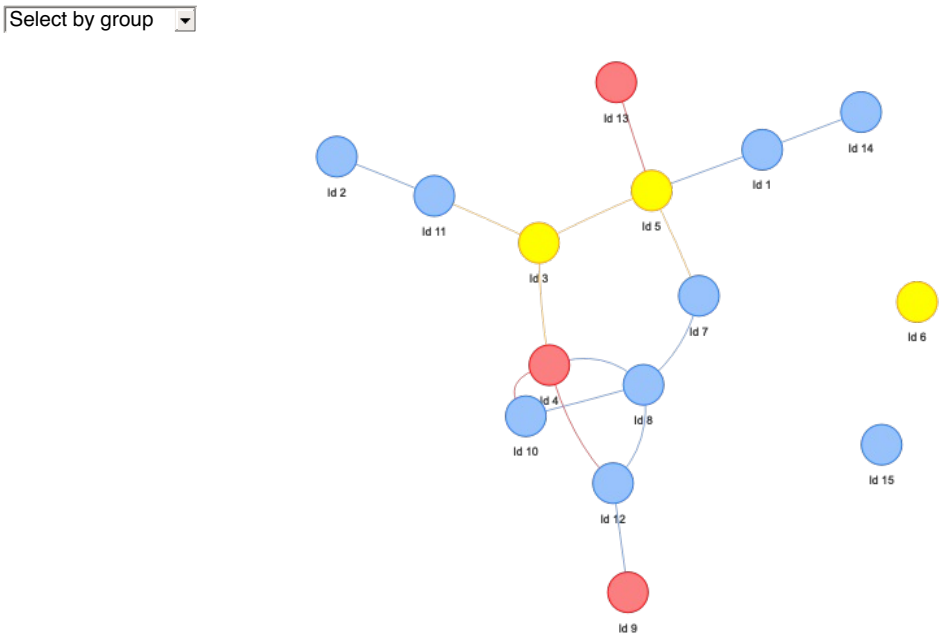



```
visNetwork(nodes, edges) %>% visOptions(highlightNearest = TRUE, nodesIdSelection = TRUE)
```

Select by id




```
visNetwork(nodes, edges) %>% visOptions(selectedBy = "group")
```



Exercice 1.8 (Interactions entre media)**.**

On considère à nouveau le graphe sur les médias

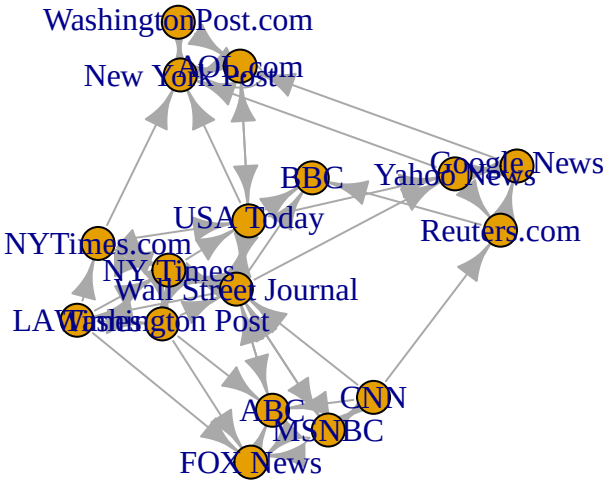
```
nodes <- read.csv("data/Dataset1-Media-Example-NODES.csv", header=T, as.is=T)
links <- read.csv("data/Dataset1-Media-Example-EDGES.csv", header=T, as.is=T)
head(nodes)
   id               media media.type type.label
1 s01            NY Times          1  Newspaper
2 s02     Washington Post          1  Newspaper
3 s03 Wall Street Journal          1  Newspaper
4 s04           USA Today          1  Newspaper
5 s05            LA Times          1  Newspaper
6 s06       New York Post          1  Newspaper
  audience.size
1            20
2            25
3            30
4            32
5            20
6            50
head(links)
  from  to weight      type
1  s01 s02     10 hyperlink
2  s01 s02     12 hyperlink
3  s01 s03     22 hyperlink
4  s01 s04     21 hyperlink
5  s04 s11     22   mention
6  s05 s15     21   mention
```

L'objet `nodes` représente les nœuds du graphe et l'objet `links` les arêtes. On définit l'objet `graphe` avec

```
media <- graph_from_data_frame(d=links, vertices=nodes, directed=T)
V(media)$name <- nodes$media
```

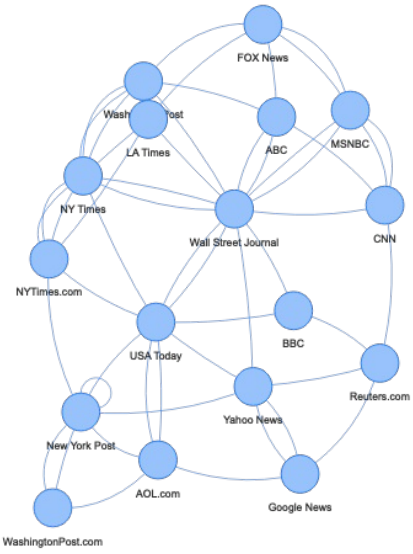
et on peut le visualiser en faisant un `plot` de cet objet

```
plot(media)
```



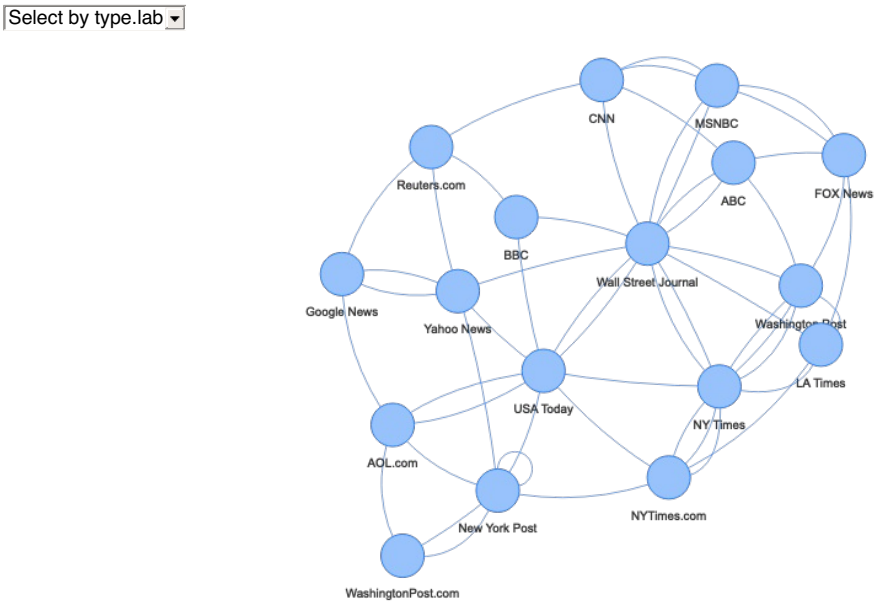
1. Visualiser ce graphe avec **VisNetwork**. On pourra utiliser la fonction **toVisNetworkData**

```
media.VN <- toVisNetworkData(media)
visNetwork(nodes=media.VN$nodes,edges=media.VN$edges)
```



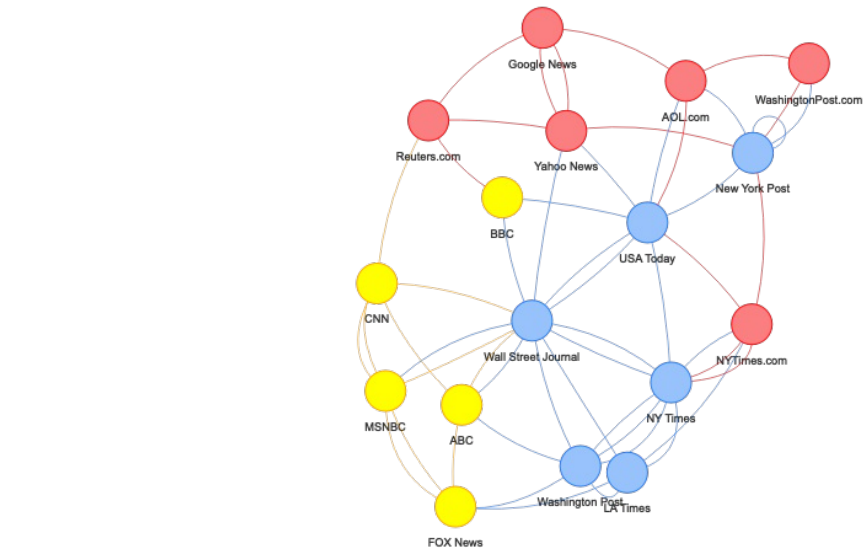
2. Ajouter une option qui permette de sélectionner le type de media (Newspaper, TV ou Online).

```
visNetwork(nodes=media.VN$nodes,edges=media.VN$edges) %>%
  visOptions(selectedBy = "type.label")
```



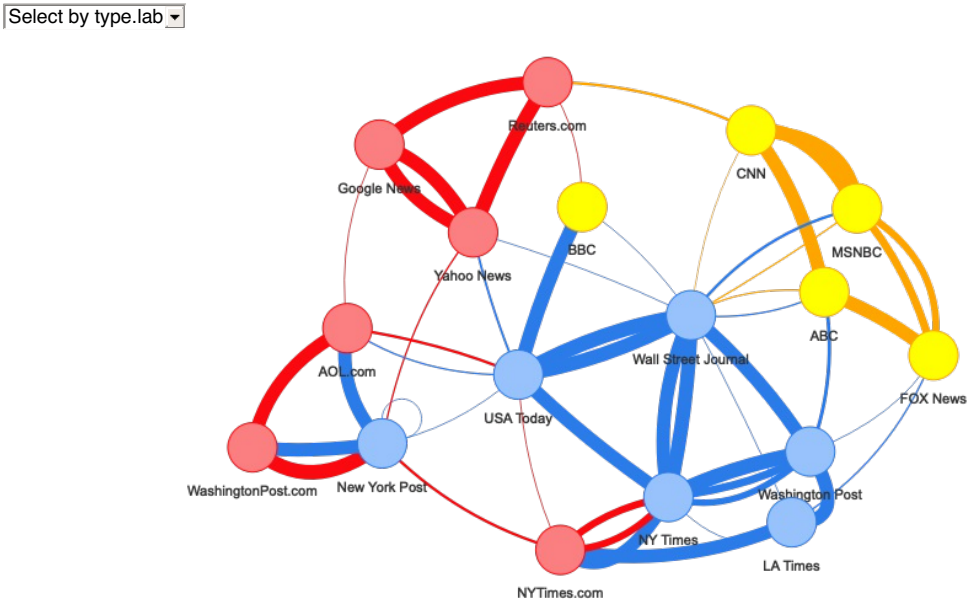
3. Utiliser une couleur différente pour chaque type de media.
   Il suffit de donner le nom ***group*** *à la variable* ***type.label****.*

```
media.VN1 <- media.VN
names(media.VN1$nodes)[3] <- "group"
visNetwork(nodes=media.VN1$nodes,edges=media.VN1$edges) %>%
  visOptions(selectedBy = "type.label")
```



4. Faire des flèches d'épaisseur différente en fonction du poids (weight). On pourra également ajouter l'option **visOptions(highlightNearest = TRUE)**.
   Il suffit de donner le nom ***value*** *à la variable weight.*

```
names(media.VN1$edges)[3] <- "value"
visNetwork(nodes=media.VN1$nodes,edges=media.VN1$edges) %>%
  visOptions(selectedBy = "type.label",highlightNearest = TRUE)
```



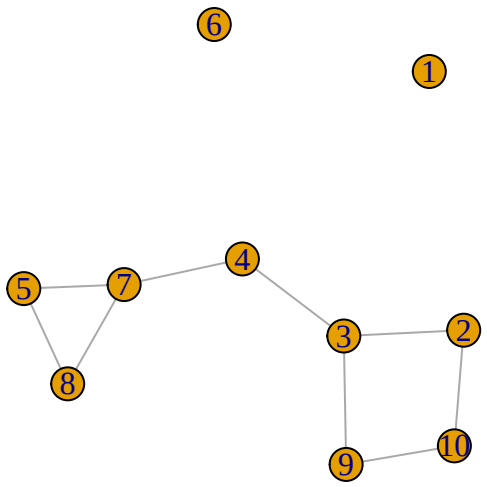
1.4.2 Graphes ggplot avec ggnet

Les fonctions **ggnet** et **ggnet2** du package `GGally` permettent de tracer des **graphes ggplot**. On pourra trouver un descriptif clair à l'url suivante https://briatte.github.io/ggnet/.

```
library(GGally)
```

On construit un premier graphe que l'on visualise avec **igraph**.

```
set.seed(1)
G <- sample_gnp(10,0.2)
plot(G)
```

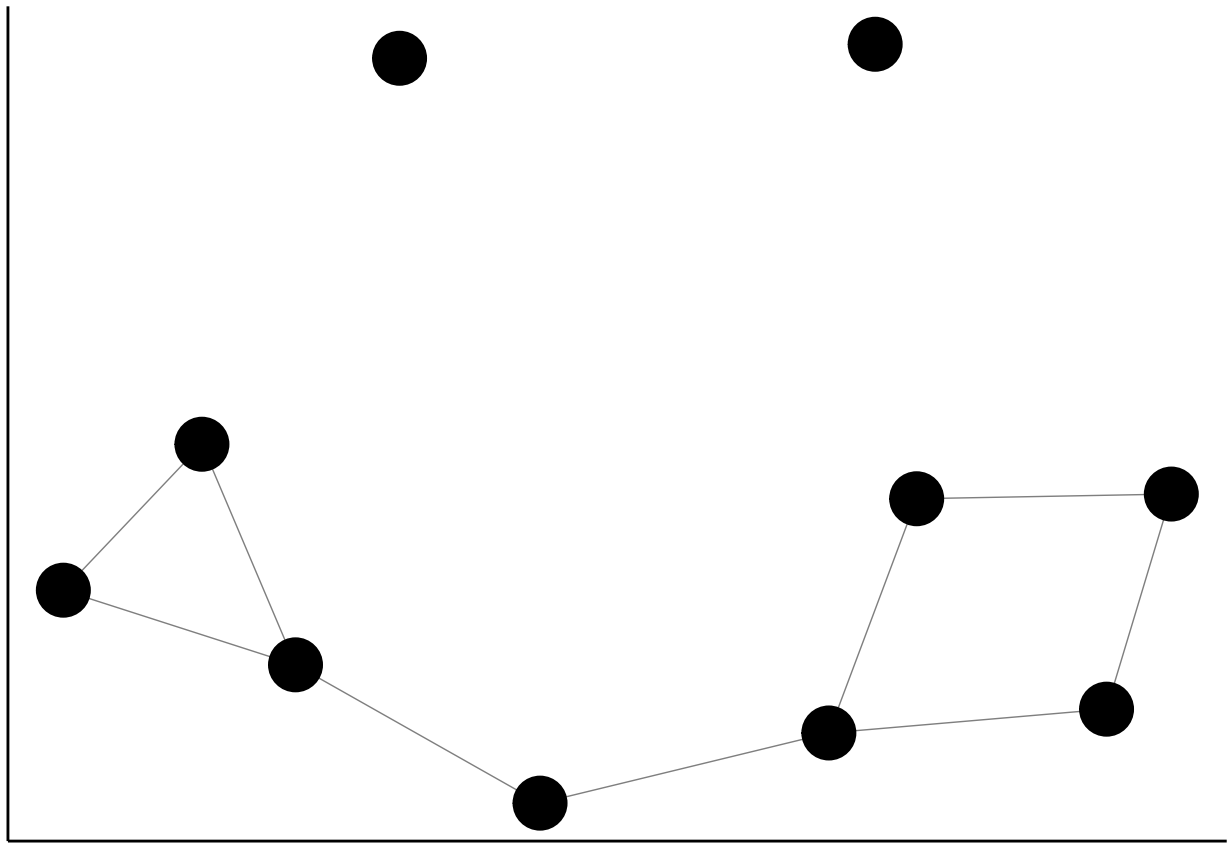


Pour visualiser ce graphe en **ggplot** il faut le transformer en objet **network** :

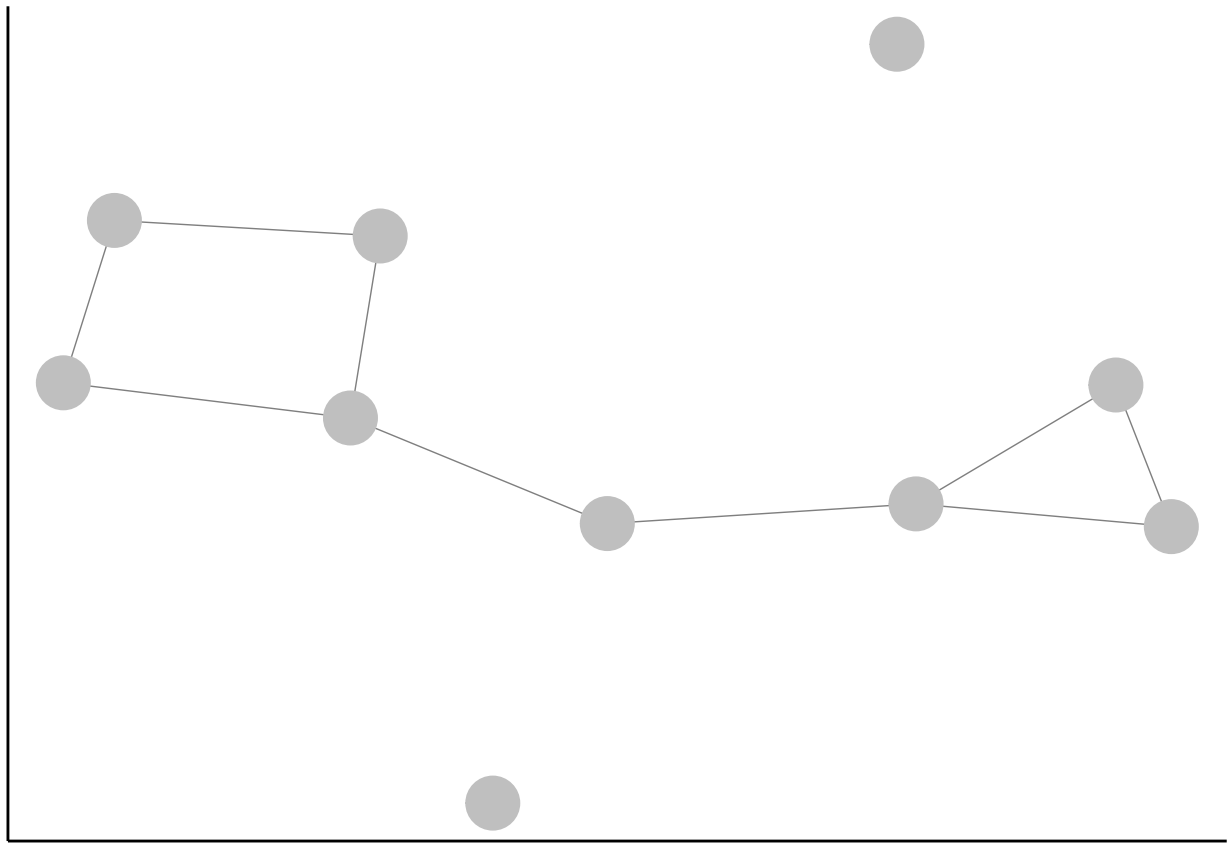
```
net <- igraph::as_data_frame(G) %>% network::as.network()
```

On peut maintenant utiliser les fonctions **ggnet** et **ggnet2**.

```
ggnet(net)
```

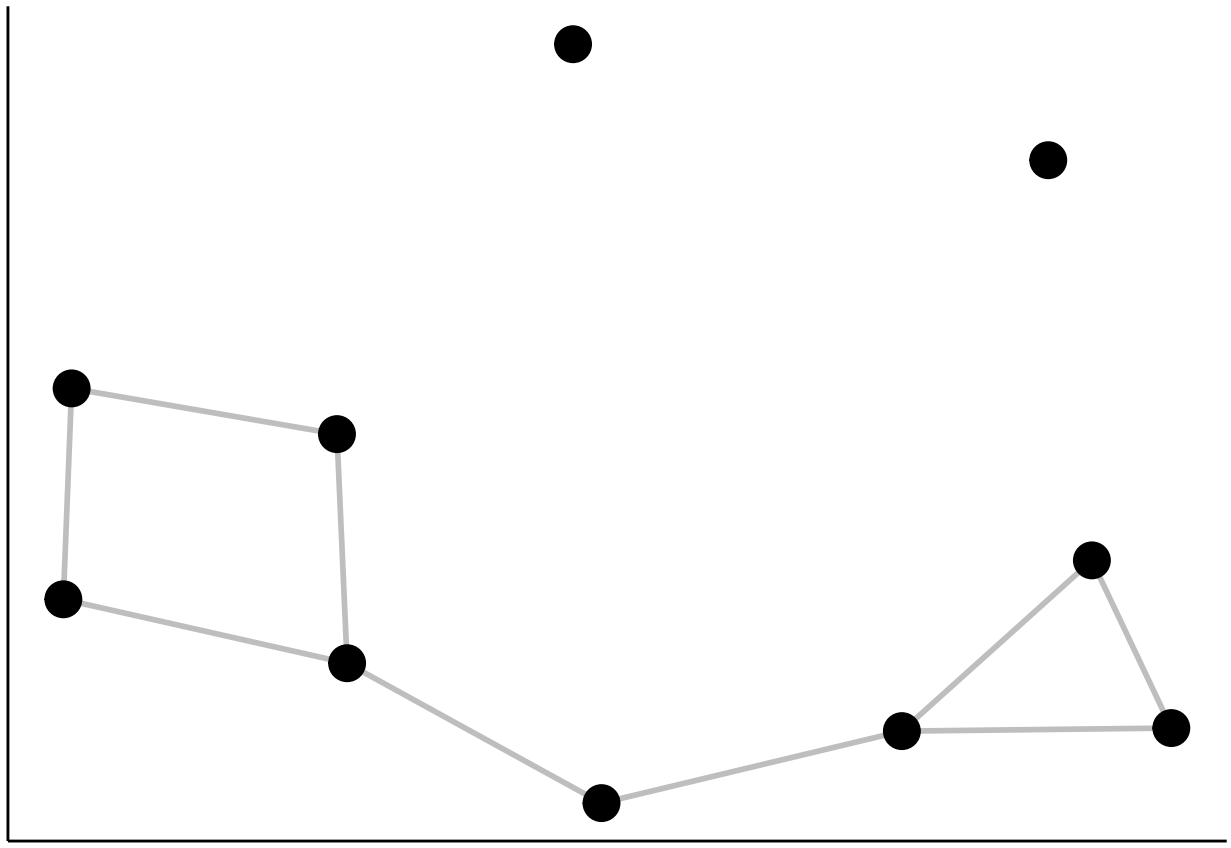


```
ggnet2(net)
```

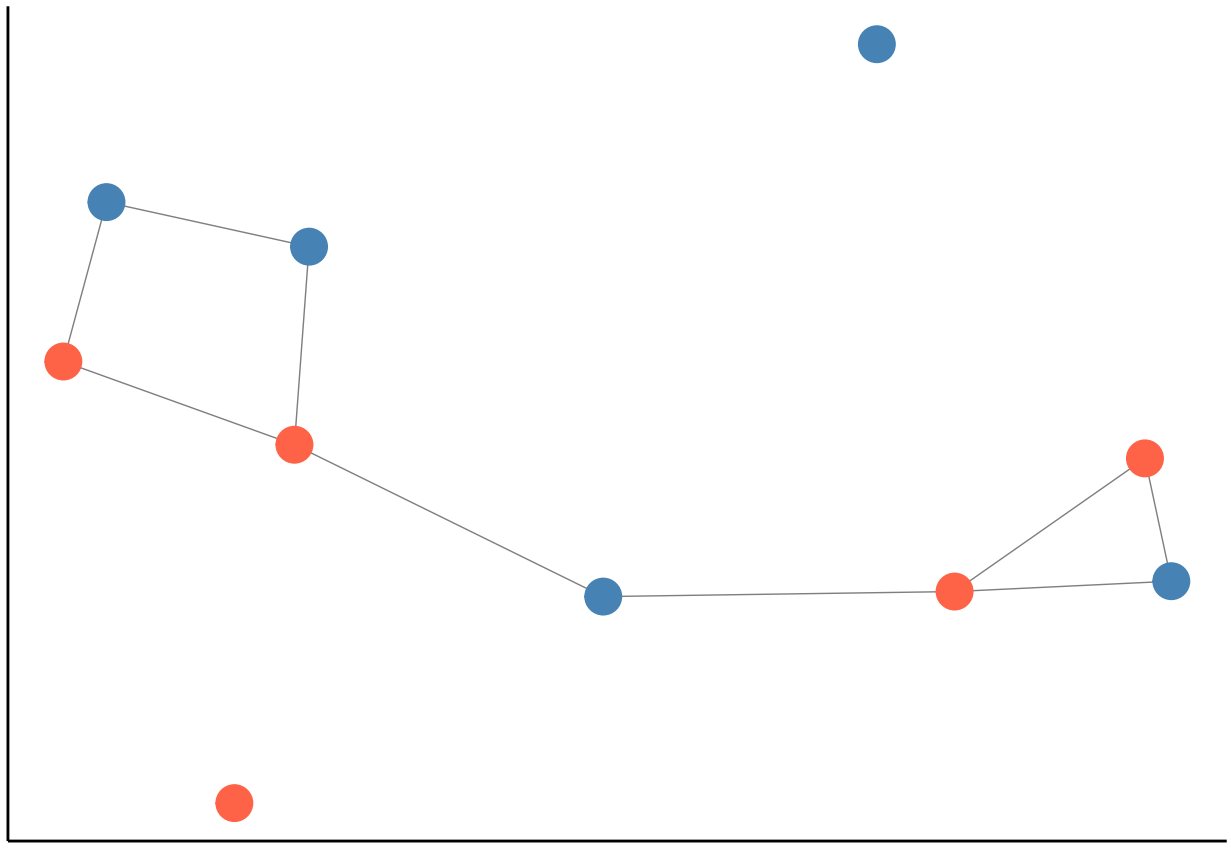


On retrouver bien entendu la plupart des options standards pour visualiser les noeuds et arêtes, par exemple

```
ggnet2(net, node.size = 6, node.color = "black", edge.size = 1, edge.color = "grey")
```



```
ggnet2(net, size = 6, color = rep(c("tomato", "steelblue"), 5))
```



2 Modèles et construction de graphes

2.1 Modèles de graphe

Nous proposons dans cette partie de générer des graphes selon différents modèles de graphes aléatoires présentés en cours.

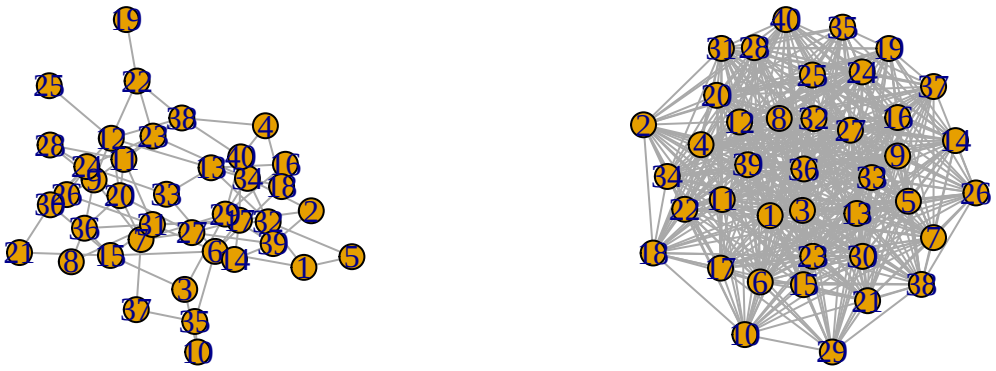
2.1.1 Graphe d'Erdos-Renyi

La fonction **sample_gnp** du package `igraph` permet de simuler un graphe d'**Erdos-Renyi**. On donne comme paramètres

- n le nombre de nœuds
- p la probabilité de connexion entre deux nœuds.

On simule 2 graphes différents : un avec peu de connexions, et un autre très connecté.

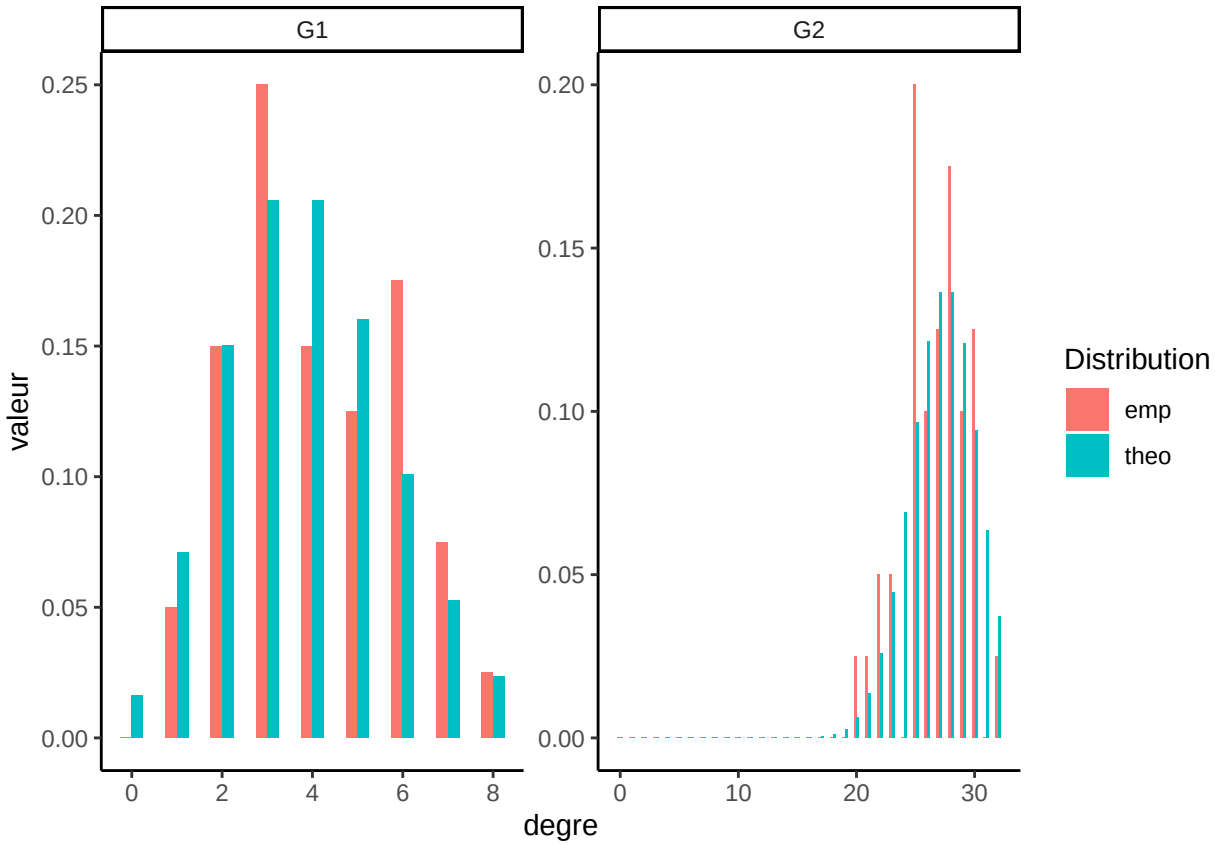
```
set.seed(1)
n <- 40
p1 <- 0.1
p2 <- 0.7
G1 <- sample_gnp(n,p1)
G2 <- sample_gnp(n,p2)
par(mfrow=c(1,2))
plot(G1)
plot(G2)
```



On rappelle que dans un graphe d'Erdos-Renyi la distribution du degrés du nœud i est binomiale : $\mathcal{B}(n-1,p)$.
Exercice 2.1 (Distribution des degrés)**.**

A l'aide d'un diagramme en barre, comparer les distributions empiriques des degrés des noeuds (on pourra utiliser **degree.distribution**) à leur distribution théorique binomiale (**dbinom**) pour les deux graphes précédents.

```
M1 <- max(degree(G1))
deg1 <- data.frame(degre=0:M1,emp=degree.distribution(G1),theo=dbinom(0:M1,n-1,p1)) %>%
  gather(key="Distribution",value="valeur",-degre)

M2 <- max(degree(G2))
deg2 <- data.frame(degre=0:M2,emp=degree.distribution(G2),theo=dbinom(0:M2,n-1,p2)) %>%
  gather(key="Distribution",value="valeur",-degre)

deg <- bind_rows(G1=deg1,G2=deg2,.id="graphe")

ggplot(deg)+aes(x=degre,y=valeur,fill=Distribution)+
  geom_bar(stat="identity",position = "dodge",width=0.5)+facet_wrap(~graphe,scales="free")+
  theme_classic()
```



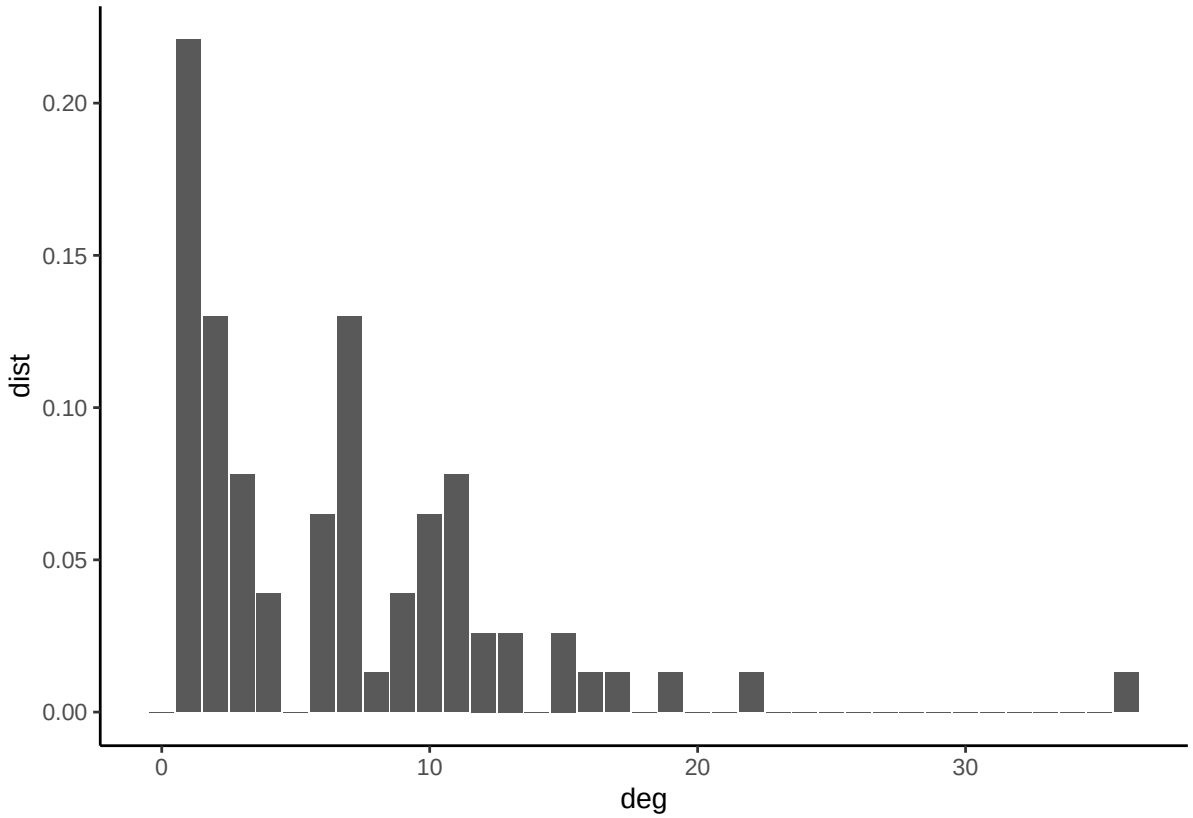
Exercice 2.2 (Les misérables sont-ils des Erdos-Renyi)**.**

On considère le graphe sur **les misérables** où une interaction entre deux personnages est définie par la co-occurrence des ces deux personnages dans un même chapitre.

```
miserab <- read.graph('data/lesmis.gml',format="gml")
```

1. Visualiser la distribution des degrés de ce graphe.

```
deg.mis <- data.frame(deg=0:max(degree(miserab)),dist=degree.distribution(miserab))
ggplot(deg.mis)+aes(x=deg,y=dist)+geom_bar(stat = "identity")+theme_classic()
```

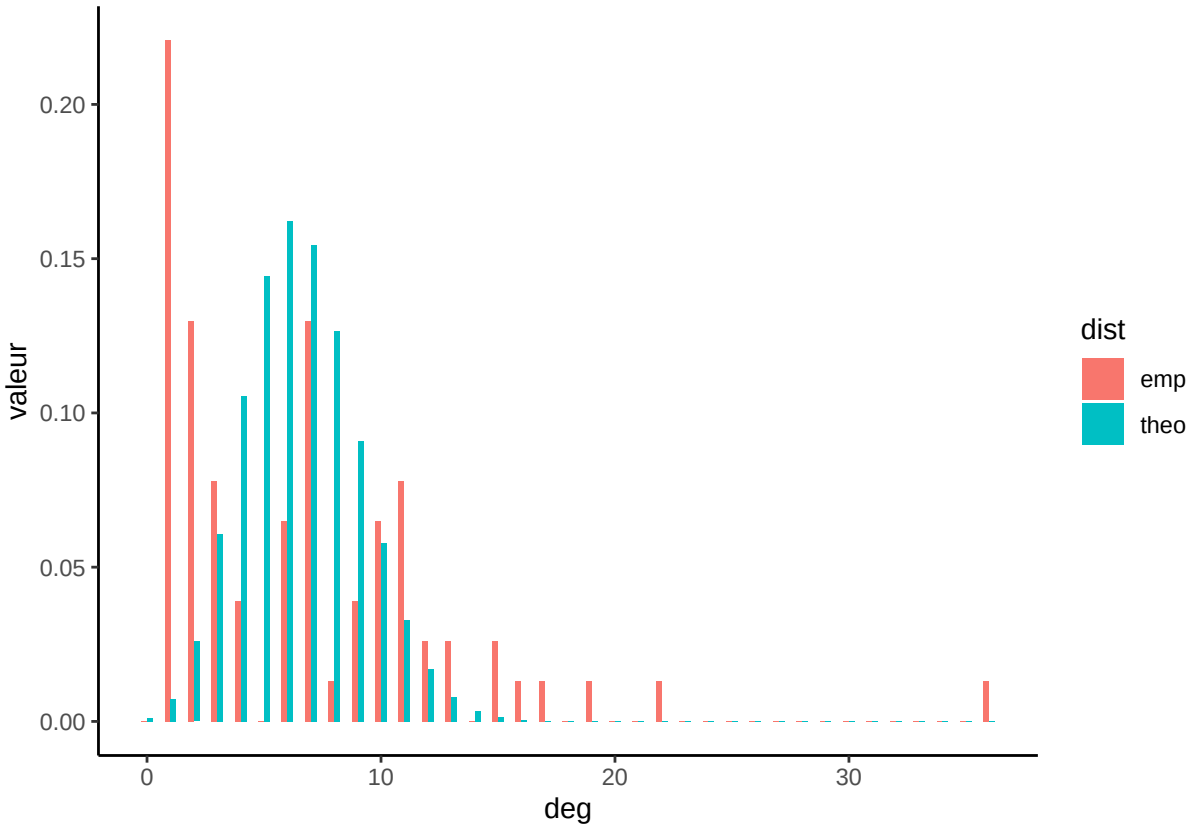


2. On souhaite comparer cette distribution à celle d'un graphe d'Erdos-Renyi. Proposer un moyen d'estimer les paramètres (n et p).
   Pour n, il suffit de prendre le nombre de noeuds. Pour p, le nombre d'arêtes observés divisé par le nombre d'arêtes possibles (C_n^2).

```
nhat <- vcount(miserab)
phat <- ecount(miserab)/choose(nhat,2)
nhat;phat
[1] 77
[1] 0.08680793
```

3. Comparer la distribution empirique du graphe à celle théorique.

```
deg.mis1 <- deg.mis %>% mutate(emp=dist,theo=dbinom(0:max(deg),nhat-1,phat)) %>% select(-dist) %>%
ggplot(deg.mis1)+aes(x=deg,y=valeur,fill=dist)+
  geom_bar(stat="identity",position = "dodge",width=0.5)
```

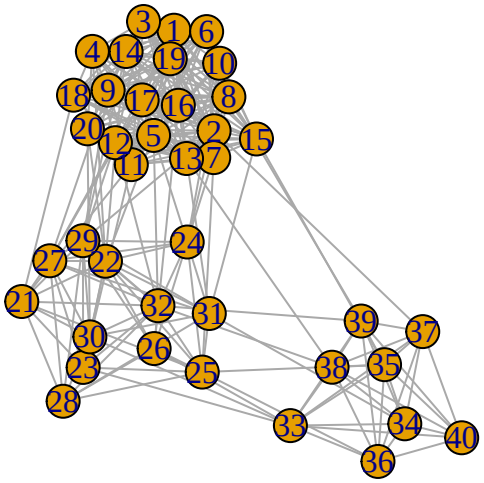


On observe que les deux distributions n'ont rien avoir l'une avec l'autre. L'hypothèse que le graphe observé est une réalisation d'un G(n,p) semble peu réaliste.

2.1.2 Modèles à blocs stochastiques

La fonction **sample_sbm** du package `igraph` permet de simuler un graphe **SBM**.

```
n <- 40 # nombre de noeuds
Q <- 3 # nombre de clusters
pi <- c(0.5, 0.3, 0.2) # appartenance aux groupes
effectifs <- n*pi

connectivite_matrix <- matrix(c(0.9, 0.1, 0.04,
                                0.1,0.7, 0.05,
                                0.04, 0.05, 0.95),nrow=Q) # matrice de connexion inter/intra groupes
G <- sample_sbm(n, pref.matrix=connectivite_matrix, block.sizes = effectifs)
plot(G)
```



On visualise qu'il s'agit bien d'un graphe avec trois **communautés** ou **groupes**, ce qui est dû aux fortes probabilités sur la diagonale de la matrice de connectivité et aux faibles valeurs de connectivité en dehors la diagonale.

Exercice 2.3 (Clustering avec un modèle **SBM**)**.**

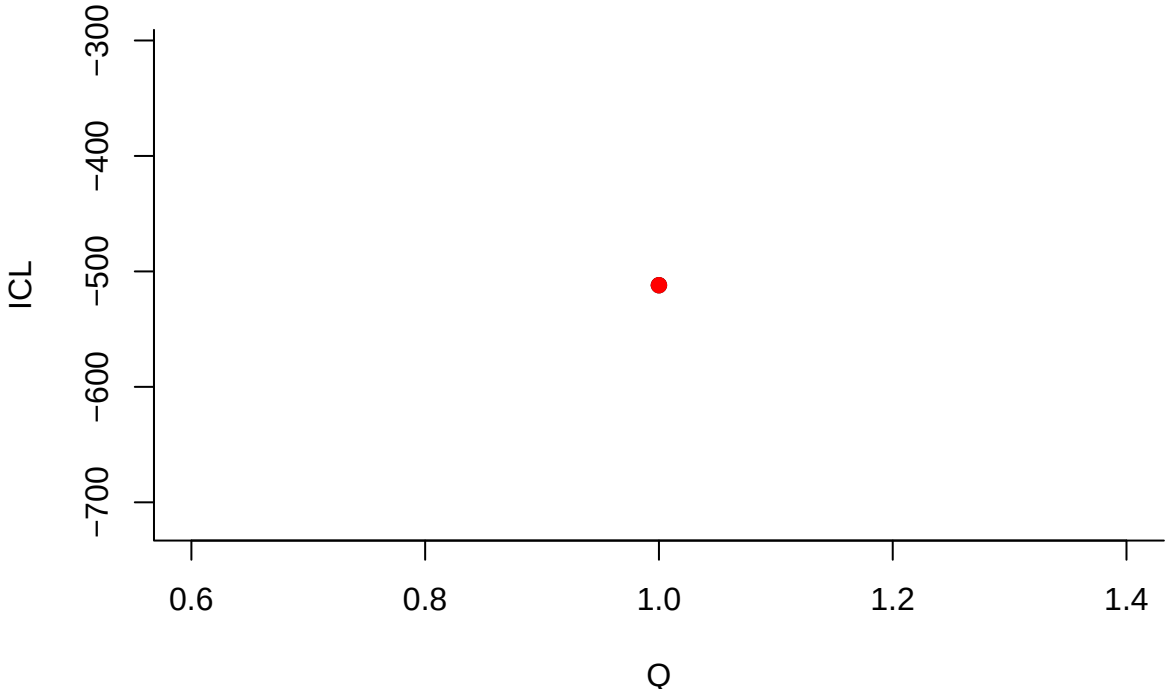
On considère le graphe G construit précédemment.

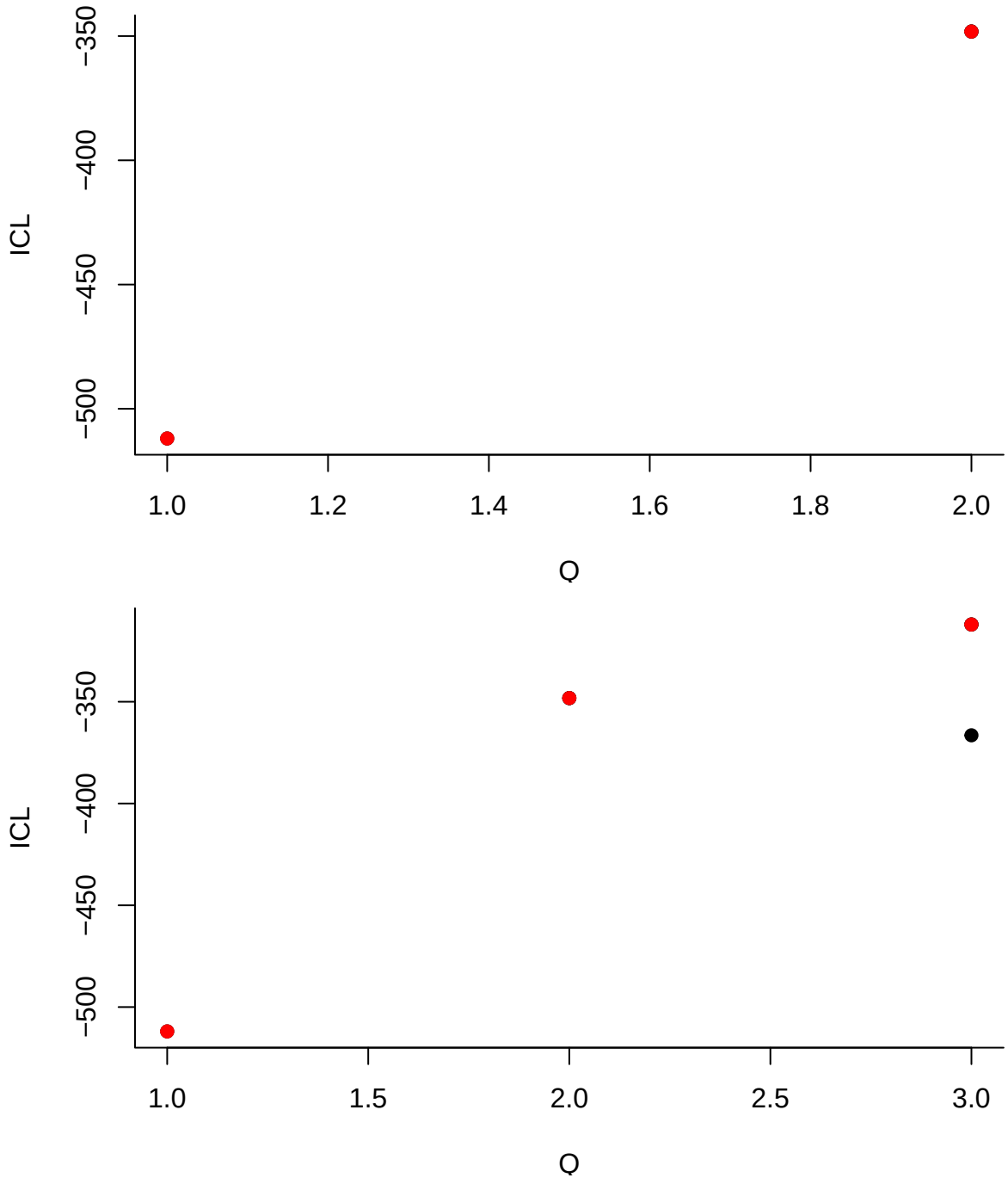
1. Calculer la matrice d'adjacence du graphe. On pourra utiliser **as_adj**.

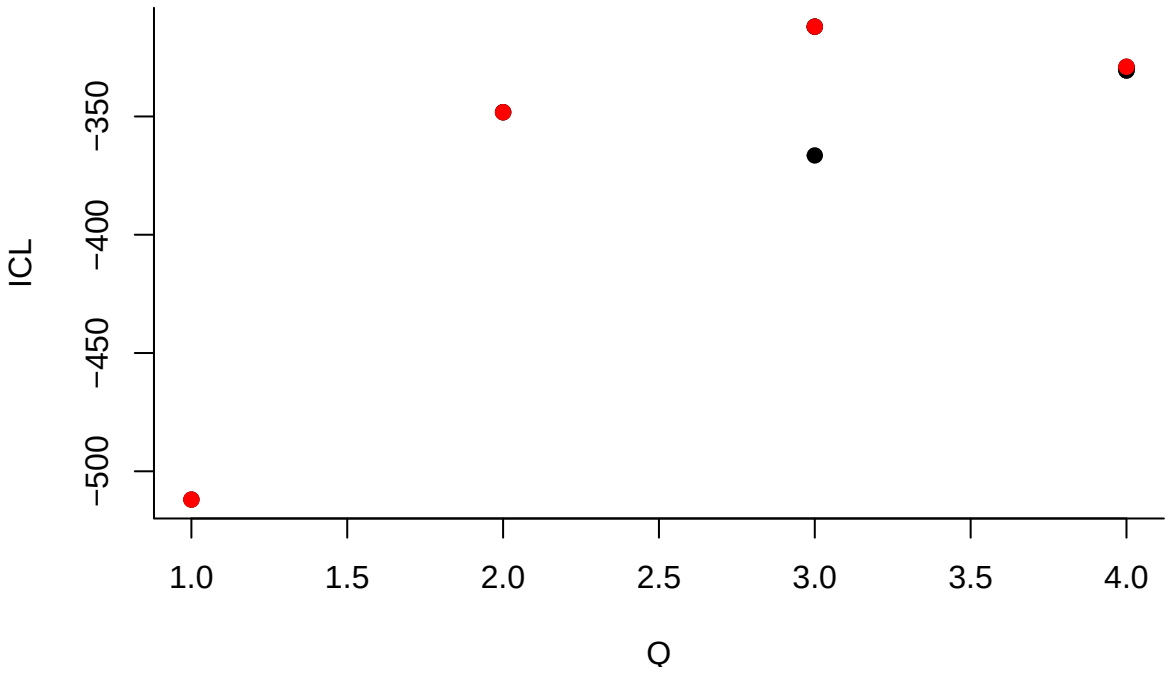
```
A <- as_adj(G, sparse=F)
```

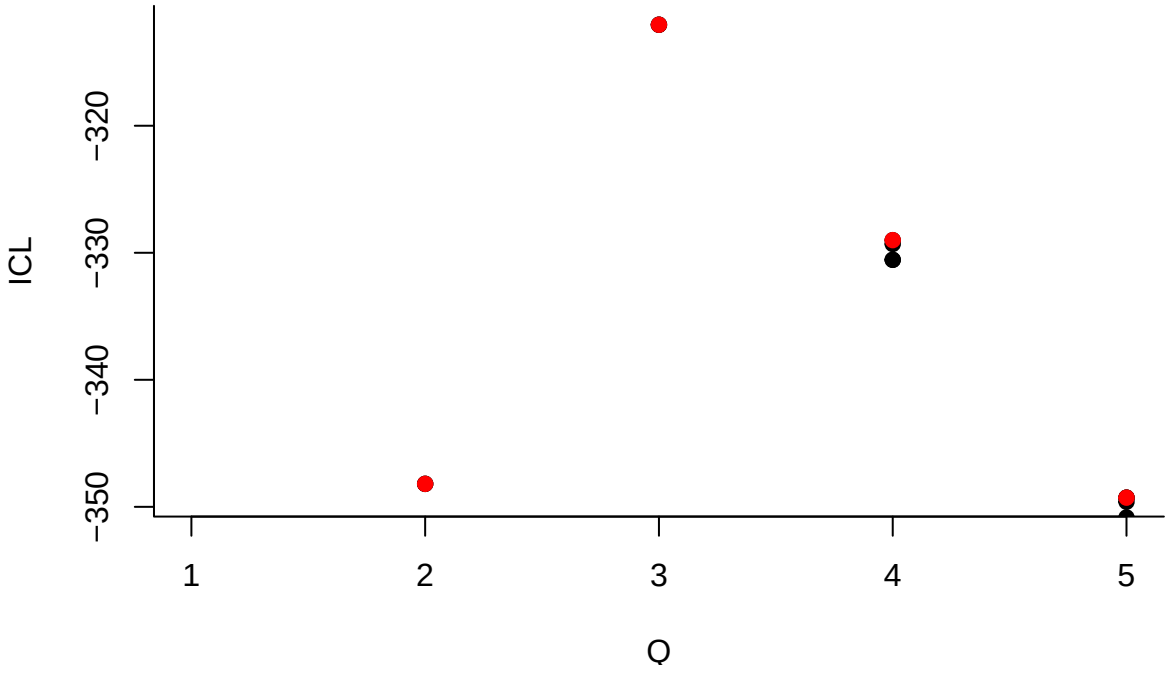
2. Les commandes suivantes permettent d'estimer les paramètres d'un graphe **SBM**

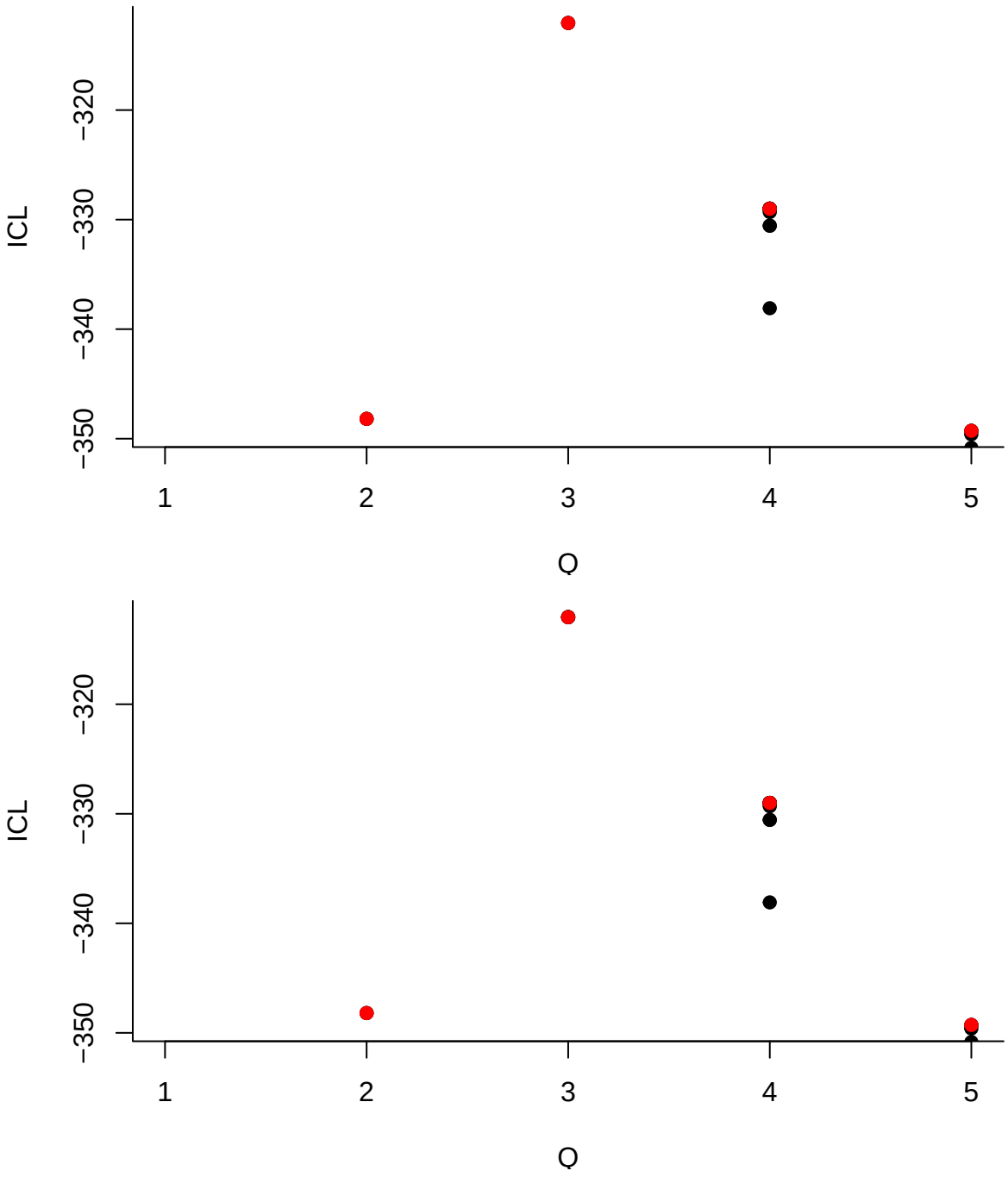
```
library("blockmodels")
mysbm <- BM_bernoulli('SBM_sym',A,verbosity=0) # SBM_sym = non dirigé
mysbm$estimate()
```

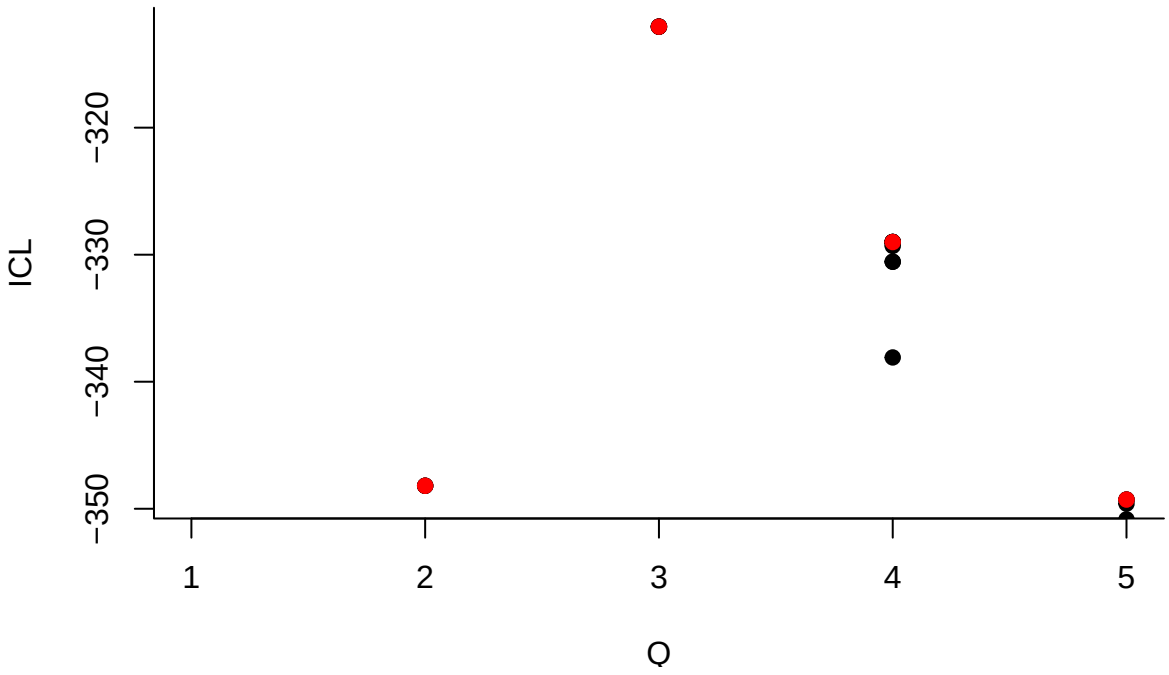



ICL
−350
−400
−450
−500
1.0
1.2
1.4
1.6
1.8
2.0
Q
ICL
−350
−400
−450
−500
1.0
1.5
2.0
2.5
3.0
Q


ICL
−350
−400
−450
−500
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Q


ICL
−320
−330
−340
−350
1
2
3
4
5
Q


ICL
−320
−330
−340
−350
1
2
3
4
5
Q
ICL
−320
−330
−340
−350
1
2
3
4
5
Q



Que pouvez-vous dire à propos du nombre de groupes ?

Le dernier graphe propose la valeur de l'ICL en fonction du nombre de clusters. On choisira donc 3 groupes.

3. À l'aide des sorties présentes dans **mysbm$model_parameters**, récupérer l'estimation de la matrice de connectivité. Comparer aux vraies valeurs.

```
mysbm$model_parameters[[3]]$pi
           [,1]       [,2]       [,3]
[1,] 0.70438063 0.10789702 0.08669494
[2,] 0.10789702 0.87471187 0.03125239
[3,] 0.08669494 0.03125239 0.97859490
```

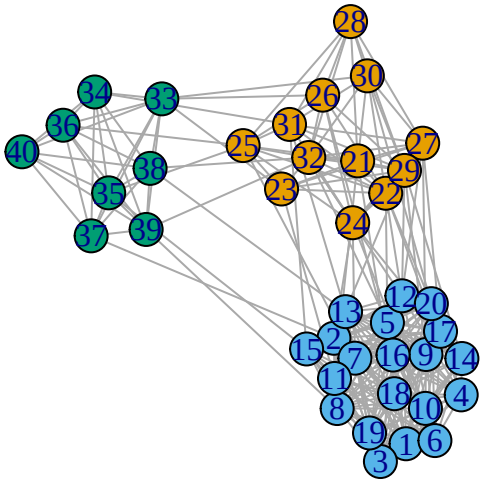
4. On trouve dans **mysbm$memberships[[3]]$Z** les estimations des probabilités a posteriori d'être dans chaque cluster. Déduire de cette matrice un groupe pour chaque observation.

```
prob.post <- mysbm$memberships[[3]]$Z
head(prob.post)
            [,1]      [,2]        [,3]
[1,] 0.002493766 0.9950125 0.002493766
[2,] 0.002493766 0.9950125 0.002493766
[3,] 0.002493766 0.9950125 0.002493766
[4,] 0.002493766 0.9950125 0.002493766
[5,] 0.002493766 0.9950125 0.002493766
[6,] 0.002493766 0.9950125 0.002493766
```

```
clust <- apply(prob.post,1,which.max) # maximum a posteriori
clust
 [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1
[29] 1 1 1 1 3 3 3 3 3 3 3 3
```

5. Visualiser les clusters sur le graphe.

```r
plot(G,vertex.color=clust)
```

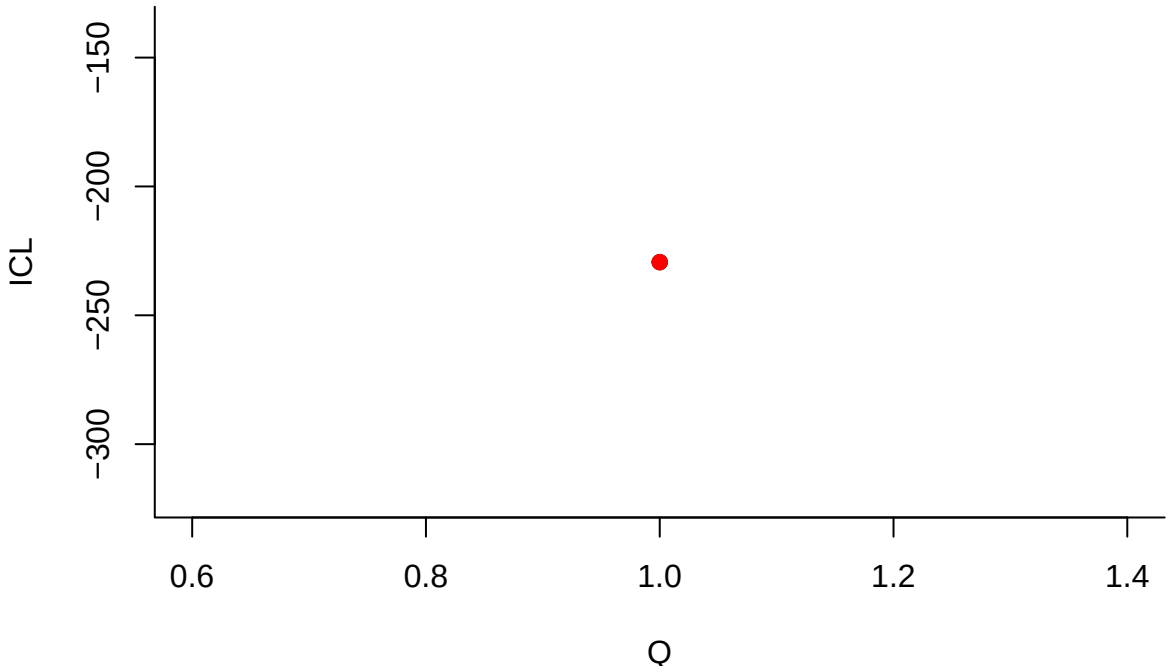


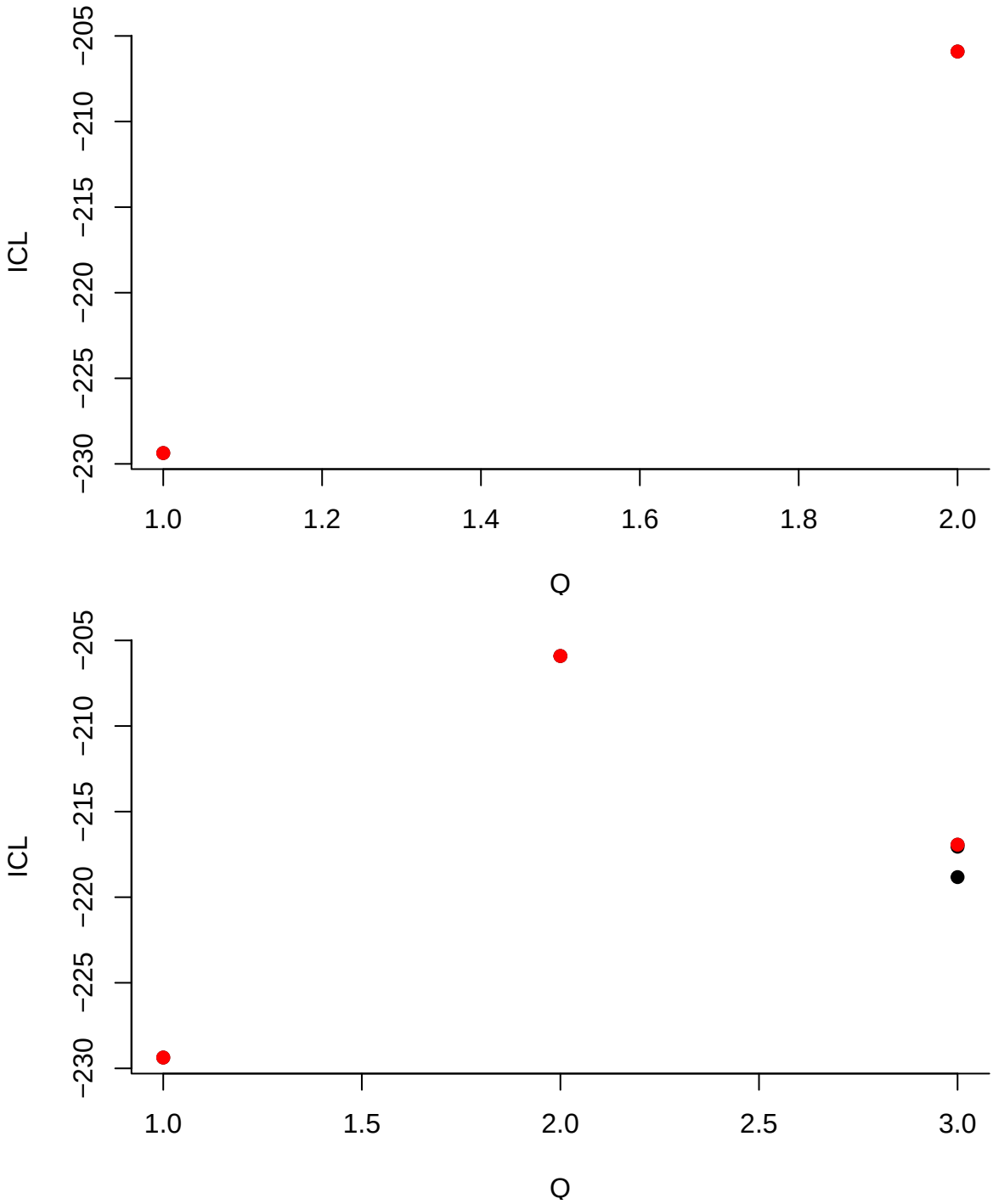
Exercice 2.4 (SBM pour le karaté)**.**

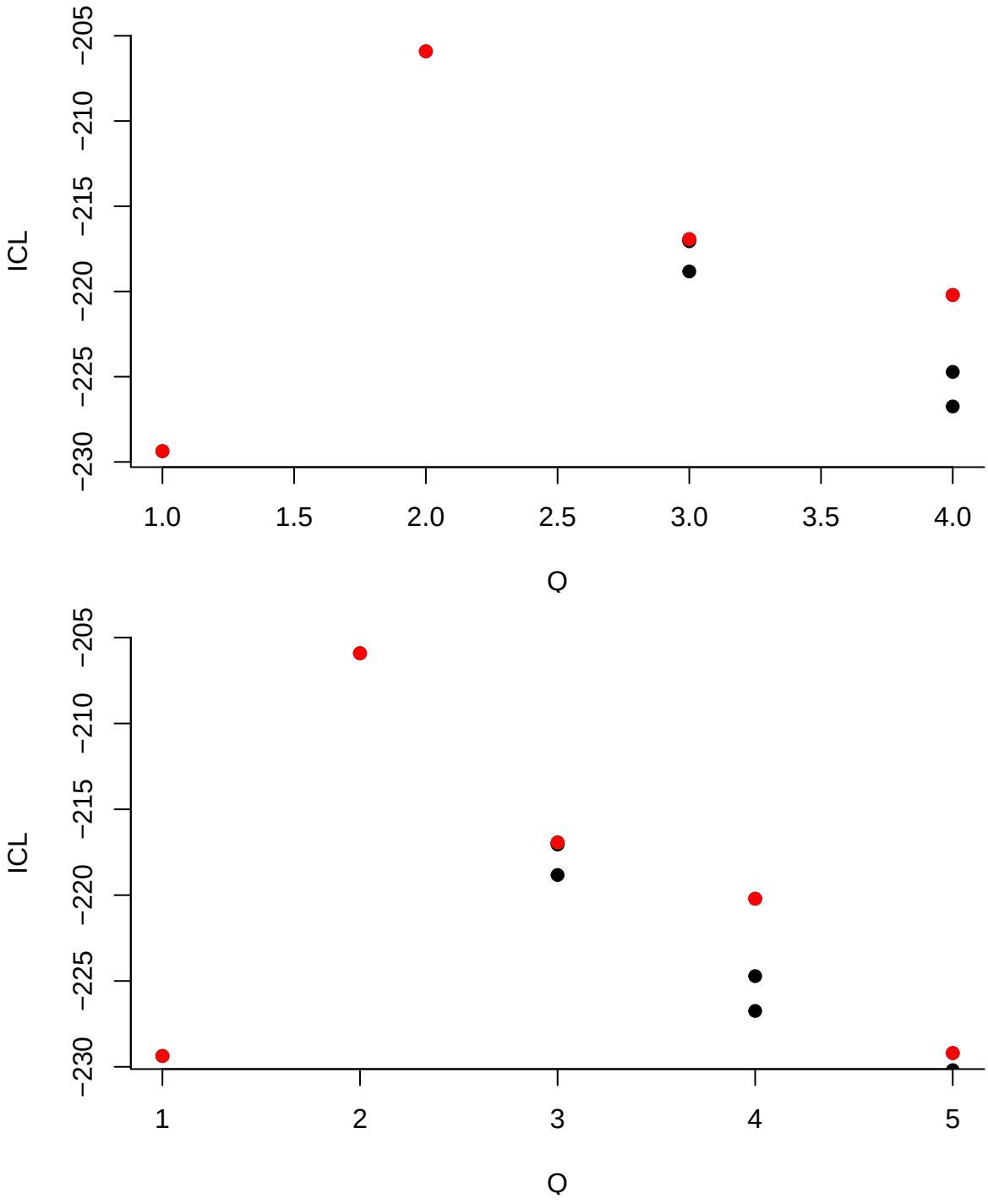
A l'aide d'un modèle SBM, identifier des clusters ou communautés sur le graphe **karate**.

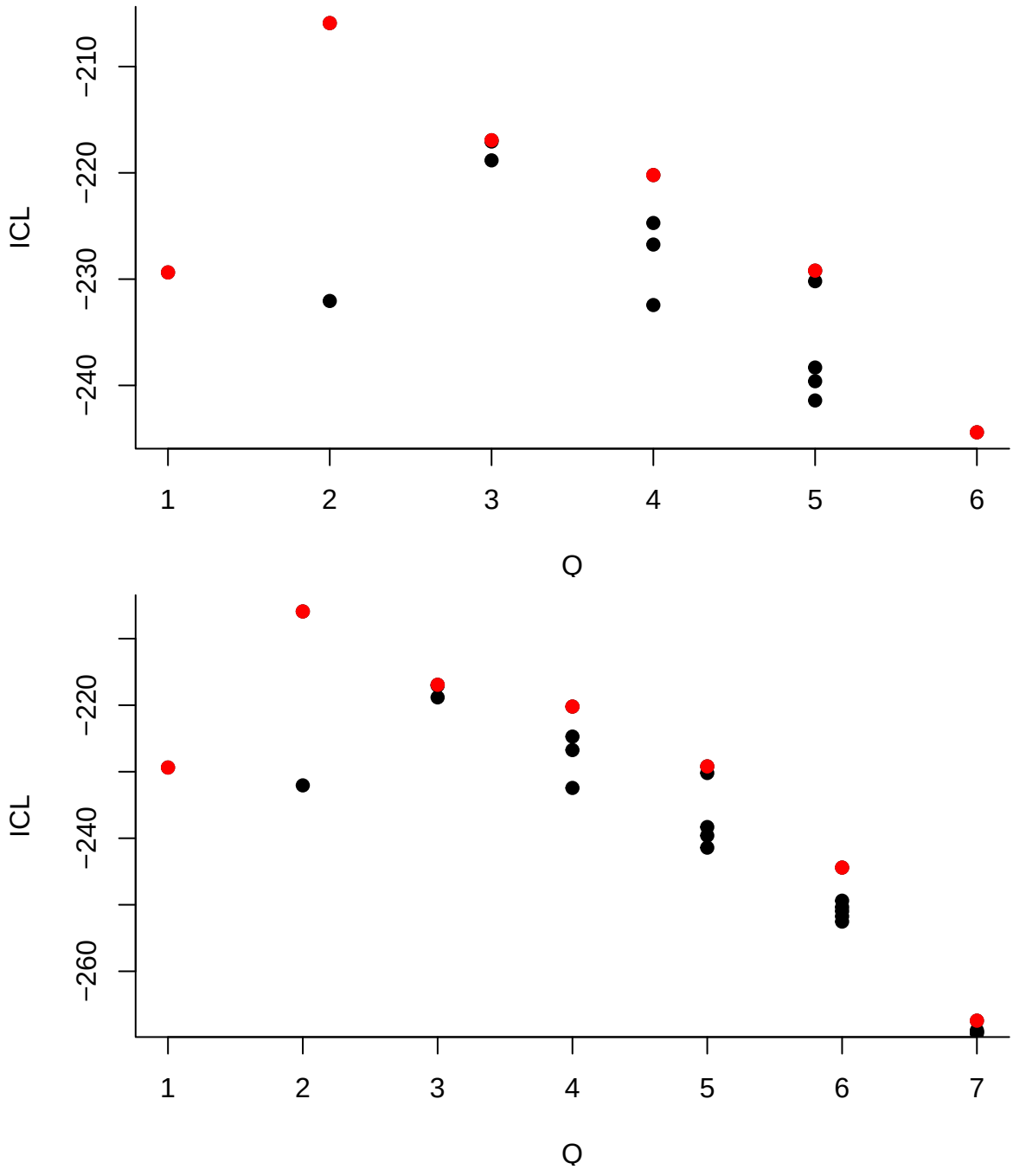
```r
library(igraphdata)
data(karate)
A <- as_adj(karate, sparse=F)
```

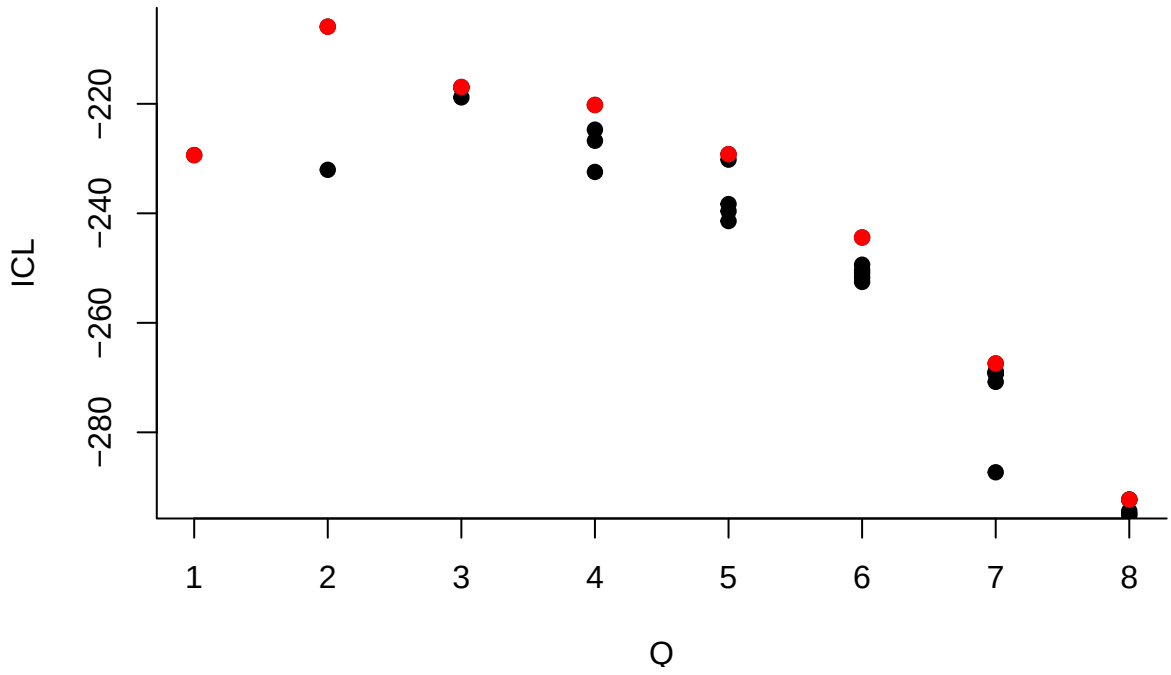
```r
sbm.kar <- BM_bernoulli('SBM_sym',A,verbosity=0,explore_min=8) # SBM_sym = non dirigé
sbm.kar$estimate()
```

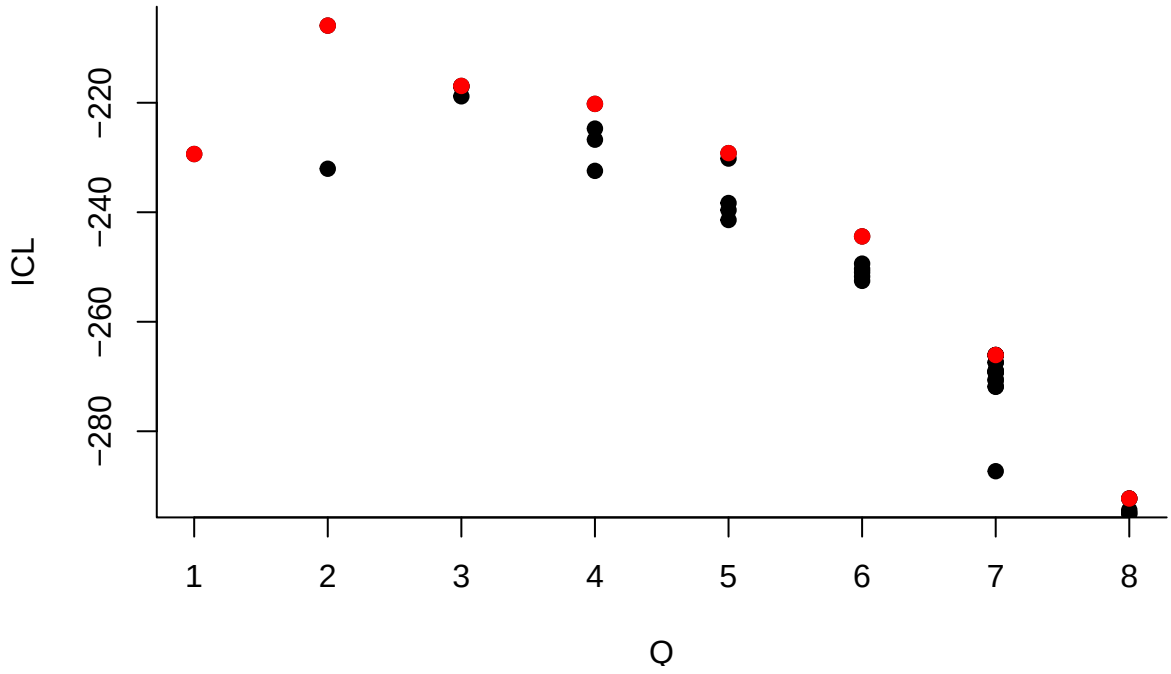


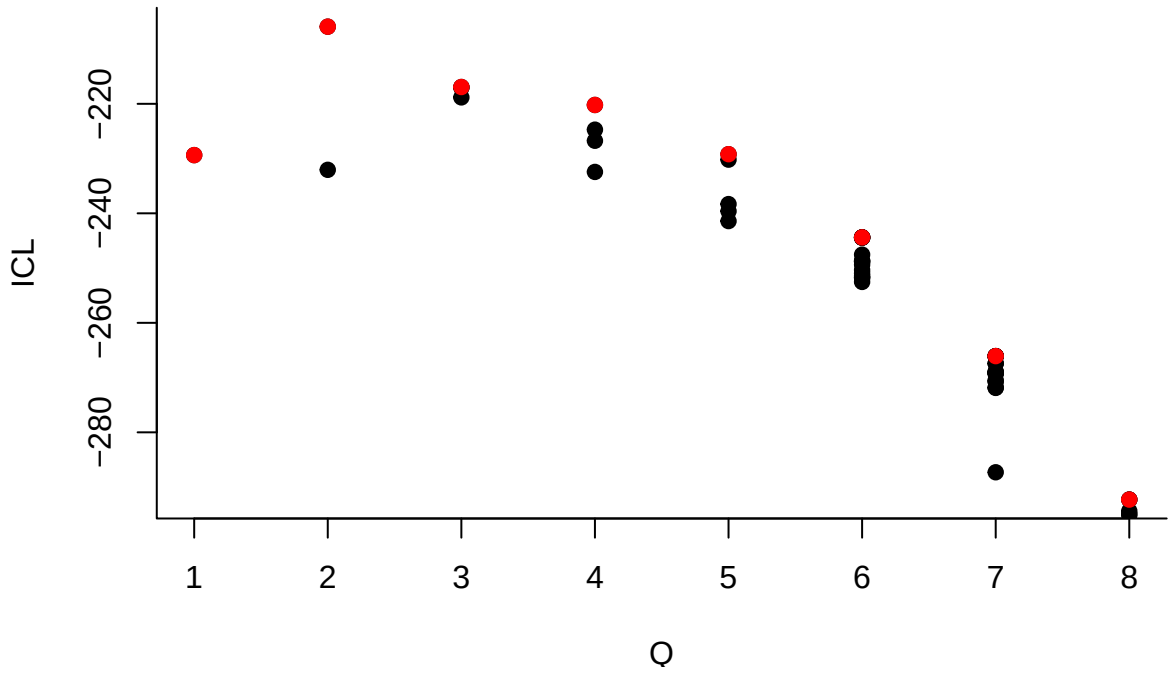

ICL
−205
−210
−215
−220
−225
−230
1.0
1.2
1.4
1.6
1.8
2.0
Q
ICL
−205
−210
−215
−220
−225
−230
1.0
1.5
2.0
2.5
3.0
Q

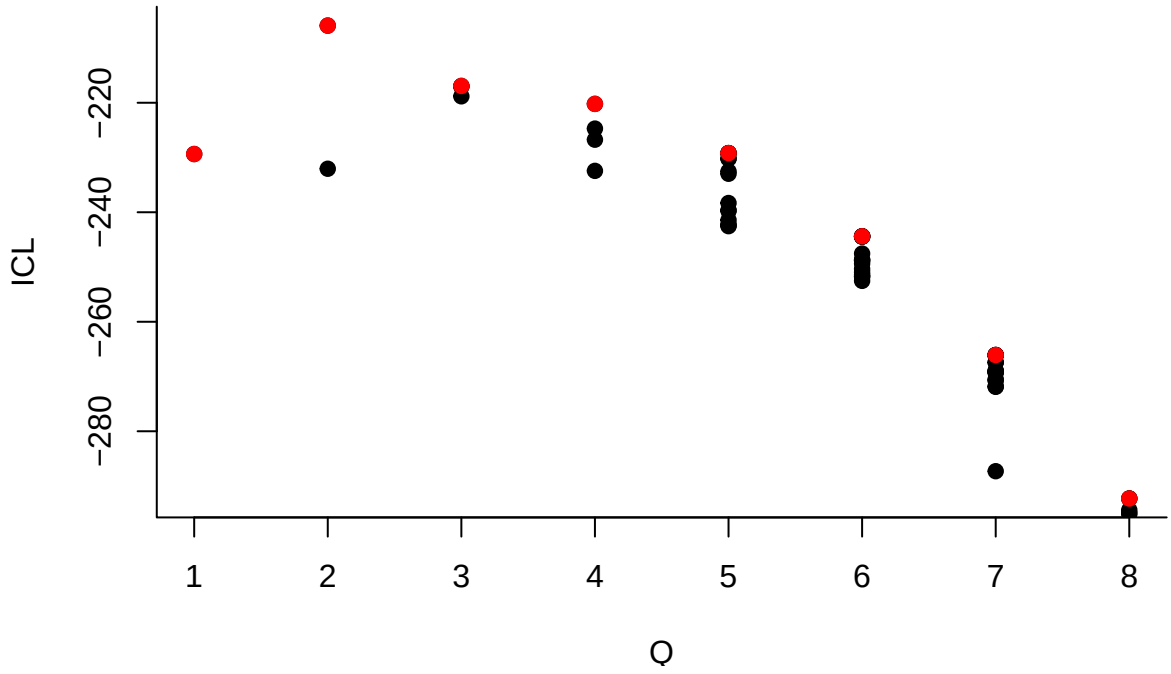

ICL
−205
−210
−215
−220
−225
−230
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Q
ICL
−205
−210
−215
−220
−225
−230
1
2
3
4
5
Q

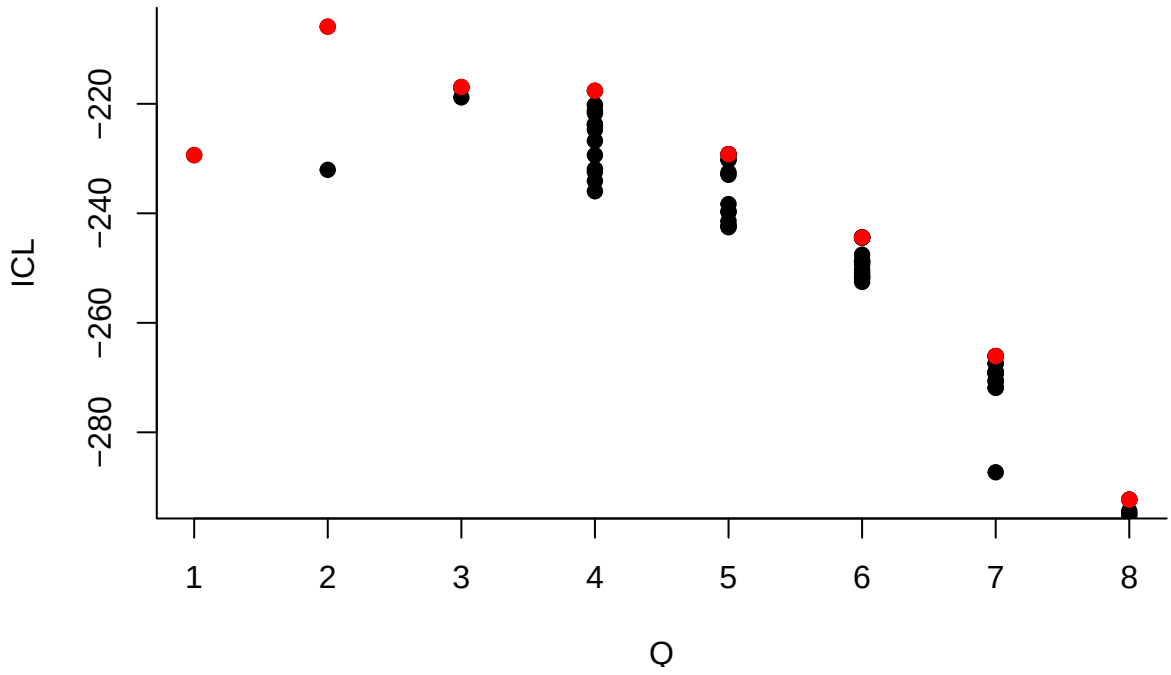

ICL
−210
−220
−230
−240
1
2
3
4
5
6
Q
ICL
−220
−240
−260
1
2
3
4
5
6
7
Q

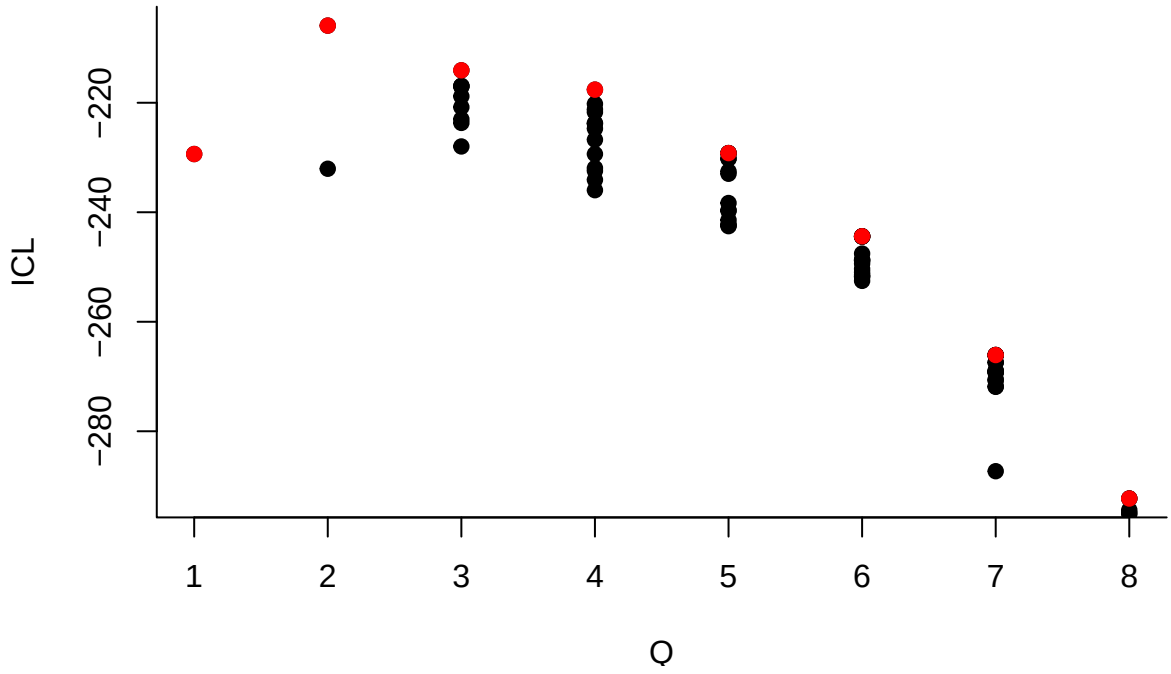

ICL
−220
−240
−260
−280
1
2
3
4
5
6
7
8
Q

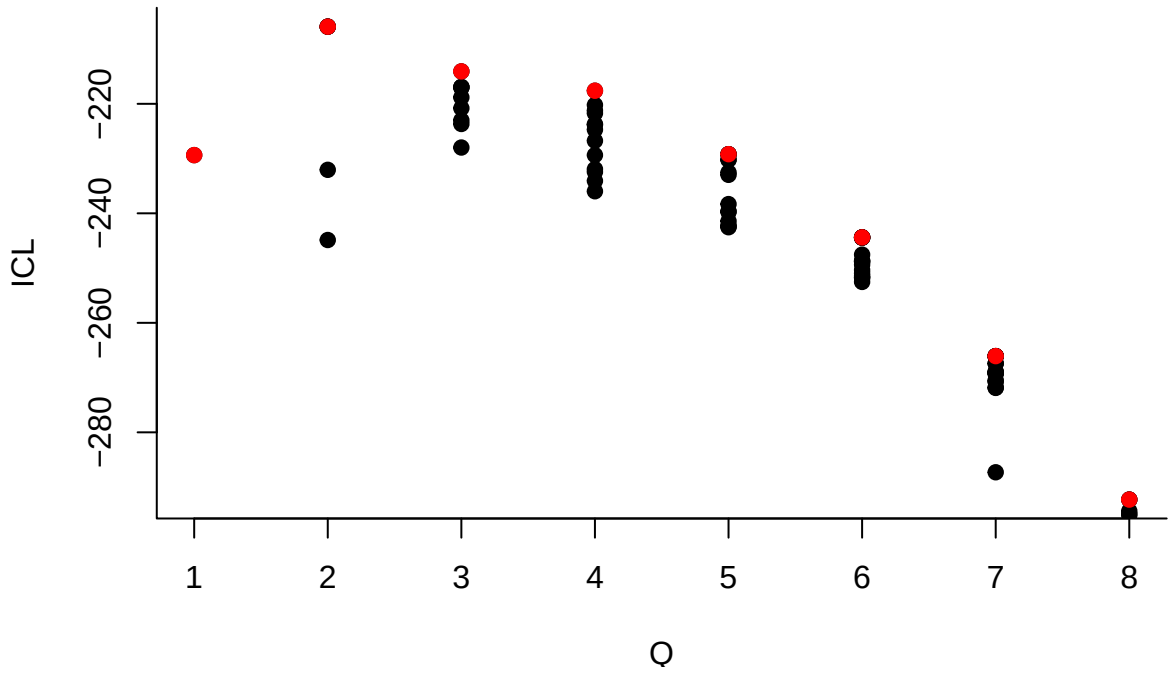

ICL
−220
−240
−260
−280
1
2
3
4
5
6
7
8
Q

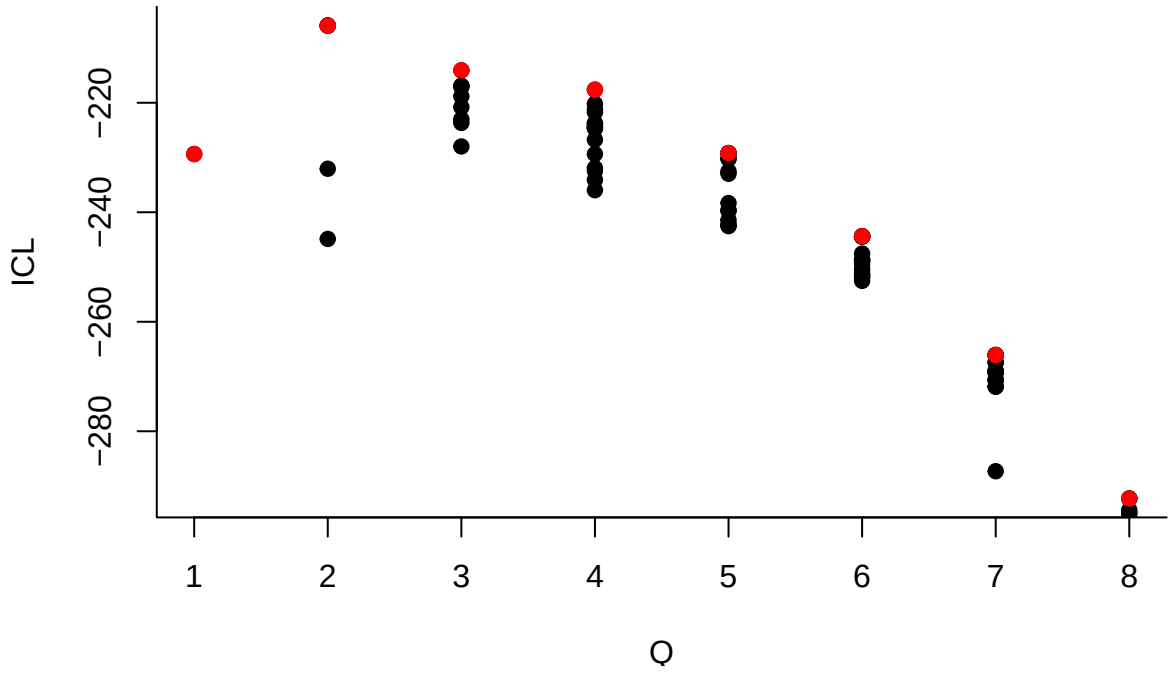

ICL
−220
−240
−260
−280
1
2
3
4
5
6
7
8
Q

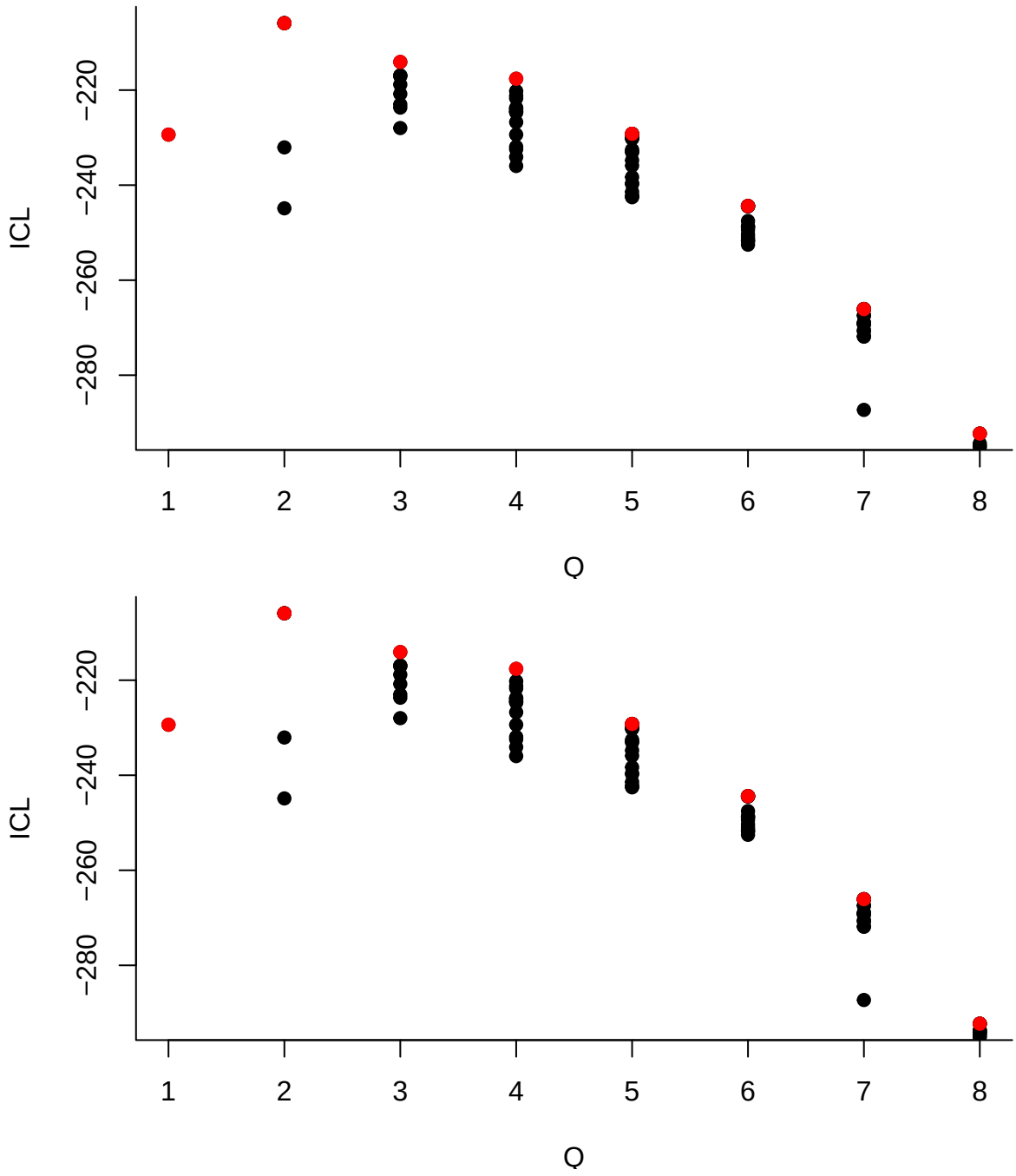

ICL
−220
−240
−260
−280
1
2
3
4
5
6
7
8
Q


ICL
−220
−240
−260
−280
1
2
3
4
5
6
7
8
Q


ICL
−220
−240
−260
−280
1
2
3
4
5
6
7
8
Q


ICL
−220
−240
−260
−280
1
2
3
4
5
6
7
8
Q


ICL
−220
−240
−260
−280
1
2
3
4
5
6
7
8
Q



On choisit 2 groupes :

```
which.max(sbm.kar$ICL)
[1] 2
```

On estime les probabilités d'appartenance à chaque groupe :

```
prob.post <- sbm.kar$memberships[[2]]$Z
head(prob.post)
             [,1]        [,2]
[1,] 0.002941176 0.997058824
```

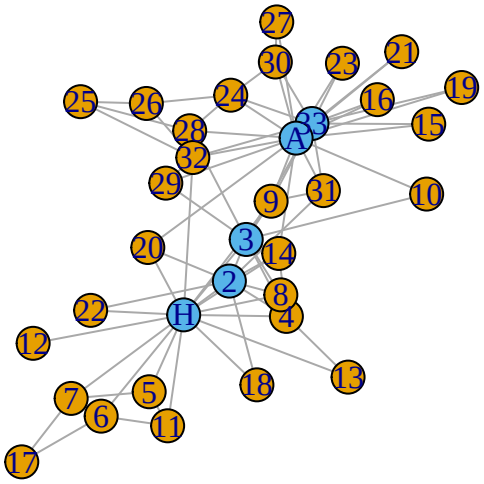
```
[2,] 0.002941176 0.997058824
[3,] 0.002941176 0.997058824
[4,] 0.997058824 0.002941176
[5,] 0.997058824 0.002941176
[6,] 0.997058824 0.002941176
```

pour en déduire un groupe pour chaque nœud

```
clust.kar <- apply(prob.post,1,which.max) # maximum a posteriori
clust.kar
 [1] 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[29] 1 1 1 1 2 2
```

que l'on visualise enfin :

```
plot(karate,vertex.color=clust.kar)
```

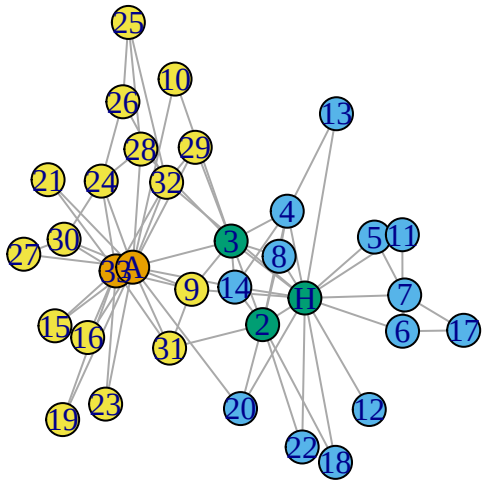


La classification n'a pas l'air d'être pertinente, on identifie un tout petit groupe de personnes très connectées. On peut essayer avec plus de groupes, par exemple 4 :

```
prob.post <- sbm.kar$memberships[[4]]$Z
head(prob.post)
            [,1]        [,2]        [,3]        [,4]
[1,] 0.002923977 0.002923977 0.991228070 0.002923977
[2,] 0.002923977 0.002923977 0.991228070 0.002923977
[3,] 0.002923977 0.002923977 0.991228070 0.002923977
[4,] 0.002923977 0.990729613 0.003422433 0.002923977
[5,] 0.002923977 0.991228070 0.002923977 0.002923977
[6,] 0.002923977 0.991228070 0.002923977 0.002923977
```

```
clust.kar <- apply(prob.post,1,which.max) # maximum a posteriori
clust.kar
 [1] 3 3 3 2 2 2 2 2 4 4 2 2 2 2 4 4 2 2 4 2 4 2 4 4 4 4 4 4 4 4 4 4 4 4
[29] 4 4 4 4 1 1
```

```
plot(karate,vertex.color=clust.kar)
```



Ça a l'air mieux !

2.2 Construire un graphe à partir de données "classiques"

Dans de nombreuses applications on ne dispose pas du graphe, l'utilisateur doit le construire à partir d'un jeu de données standard **individus-variabes** Les méthodes classiques consistent à calculer des distances entre les individus et à mettre une arrête lorsque des individus sont "proches". La notion de proximité est bien entendu à définir, il existe plusieurs possibilités :

- ε-neighborhood graph : on met une arête entre i et j si la distance entre i et j est plus petite qu'un seuil ε ;
- plus proches voisins : on met une arête entre i et j si i est parmi les plus proches voisins de j.

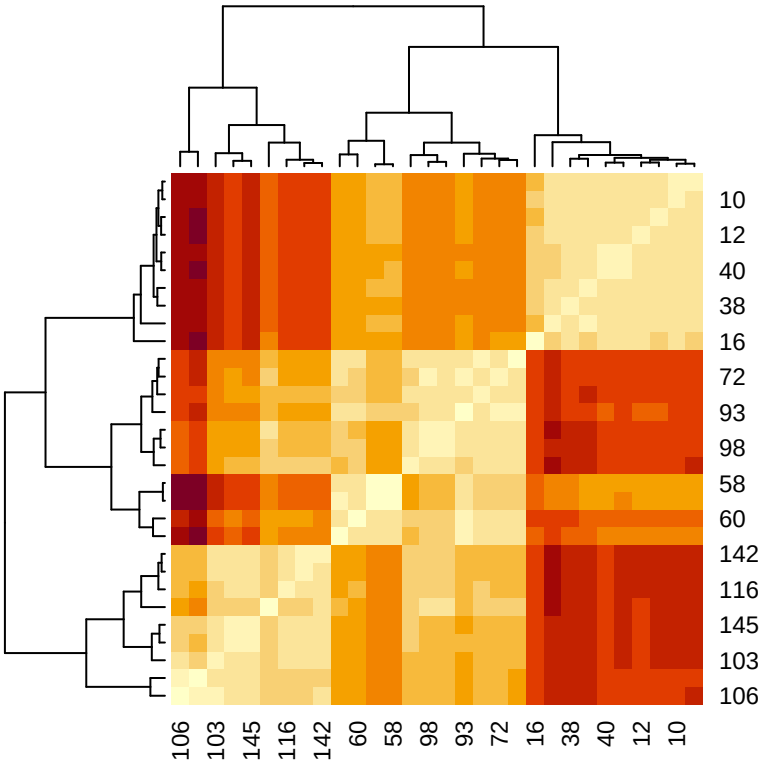
Exercice 2.5 (Plus proches voisins pour les iris)**.**

On reprend le jeu de données **iris** vu en cours, dont on extrait un sous échantillon.

```
data(iris)
set.seed(12345)
donnees <- iris[sample(nrow(iris),30),]
head(donnees)
    Sepal.Length Sepal.Width Petal.Length Petal.Width
142          6.9         3.1          5.1         2.3
51           7.0         3.2          4.7         1.4
58           4.9         2.4          3.3         1.0
93           5.8         2.6          4.0         1.2
75           6.4         2.9          4.3         1.3
96           5.7         3.0          4.2         1.2
       Species
142  virginica
51  versicolor
58  versicolor
93  versicolor
75  versicolor
96  versicolor
```

1. Construire les distances euclidiennes entre individus en ne considèrant que les 4 variables quantitatives. On stockera ces distances dans une matrice et on visualisera cette matrice à l'aide d'un **heatmap**.

```
D <- as.matrix(dist(donnees[,-5]))
heatmap(D)
```



2. A l'aide de la fonction **nng** du package `cccd`, construire :
   - un graphe de plus proches voisins à 20 ppv
   - un graphe de plus proches voisins à 2 ppv
   - un graphe de plus proches mutuels voisins à 20 ppv
   - un graphe de plus proches mutuels voisins à 2 ppv

   Ces 4 graphes seront non dirigés.

```
library(cccd)
G2 <- as.undirected(nng(dx=D,k=2,mutual=FALSE))
G20 <- as.undirected(nng(dx=D,k=20,mutual=FALSE))
GM2 <- as.undirected(nng(dx=D,k=2,mutual=TRUE))
GM20 <- as.undirected(nng(dx=D,k=20,mutual=TRUE))
```

3. Comparer les nombres d'arêtes de chaque graphe.

```
ecount(G2)
[1] 44
ecount(G20)
[1] 354
ecount(GM2)
[1] 17
ecount(GM20)
[1] 246
```

On remarque que :

- *le nombre d'arêtes augmente lorsqu'on augmente le nombre de voisins (normal).*
- *les ppv mutuels ont moins d'arêtes (normal aussi, ce critère est plus exigeant pour définir une arête).*

4. On considère maintenant le graphe de ppv (non mutuels) à 10 ppv. Ajuster un modèle SBM à 3 groupes sur ce graphe. Comparer les groupes obtenus aux espèces d'iris.

On commence par construire le graphe

```
G10 <- as.undirected(nng(dx=D,k=10,mutual=FALSE))
```

Puis on ajuste le modèle SBM :

```
A <- as_adj(G10, sparse=F)
sbm.iris <- BM_bernoulli('SBM_sym',A,verbosity=0,explore_min=8,plotting="") # SBM_sym = non dirigé
sbm.iris$estimate()
```

L'ICL est maximum pour 4 groupes

```
which.max(sbm.iris$ICL)
[1] 4
```

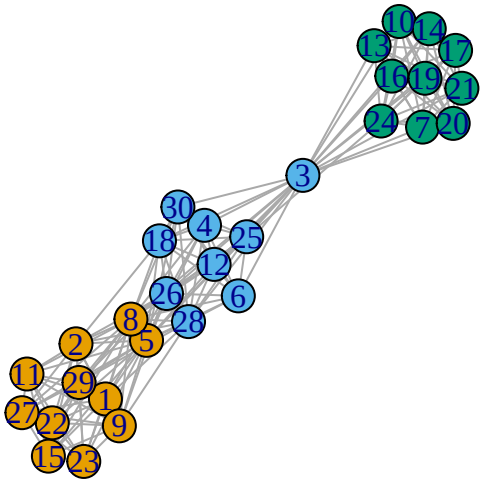
mais on en choisit 3 comme indiqué :

```
prob.post <- sbm.iris$memberships[[3]]$Z
head(prob.post)
            [,1]        [,2]        [,3]
[1,] 0.993355482 0.003322259 0.003322259
[2,] 0.993355482 0.003322259 0.003322259
[3,] 0.003322259 0.993355482 0.003322259
[4,] 0.003322259 0.993355482 0.003322259
[5,] 0.987674040 0.009003701 0.003322259
[6,] 0.003322259 0.993355482 0.003322259
```

```
clust.iris <- apply(prob.post,1,which.max) # maximum a posteriori
clust.iris
 [1] 1 1 2 2 1 2 3 1 1 3 1 2 3 3 1 3 3 2 3 3 3 1 1 3 2 2 1 2
[29] 1 2
table(clust.iris,donnees$Species)

clust.iris setosa versicolor virginica
         1      0          3         8
         2      0          9         0
         3     10          0         0
```

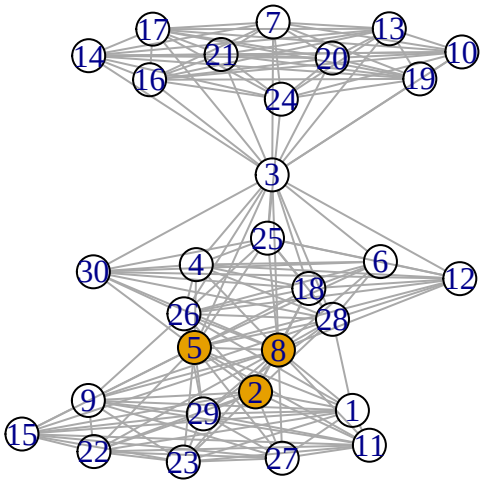
On remarque que, à 3 exceptions près, les clusters correspondent aux espèces. On peut ainsi visualiser les groupes

```
plot(G10,vertex.color=clust.iris)
```



et les 3 nœuds "mal classés" :

```
df <- tibble(clust=clust.iris,Species=donnees$Species) %>%
  mutate(clust1=fct_recode(as.character(clust),virginica="1",versicolor="2",
                           setosa="3"),err=(clust1!=Species))
plot(G10,vertex.color=df$err)
```



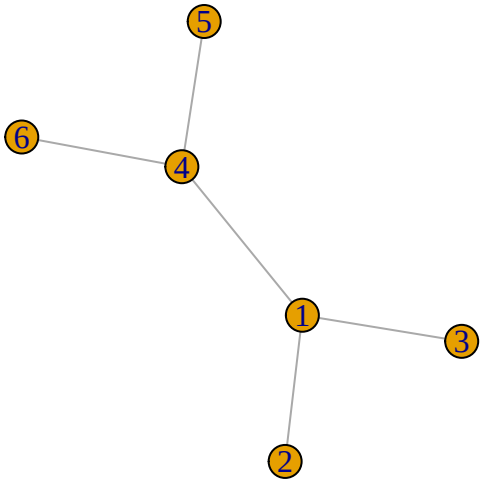
3 Détection de communautés : approche modularité

Une problématique souvent liée aux graphes est la **détection de communautés**. Elle consiste à trouver des groupes de nœuds très liés entre eux. Cette thématique est proche du clustering. Nous présentons dans cette partie les approches liés à la **modularité**. Cette dernière est un critère qui permet de mesurer la performance d'une partition de nœuds dans un graphe, plus la modularité est grande, meilleure est la partition.

Exercice 3.1 (Calculs de modularité)**.**

On considère le graphe suivant

```
G <- make_graph(c(1,2,1,3,1,4,4,5,4,6),directed = FALSE)
plot(G)
```



et les deux partitions des nœuds suivantes.

```
cl1 <- c(1,1,1,2,2,2)
cl2 <- c(1,2,1,2,1,1)
```

1. Calculer la **modularité** pour ces deux partitions en utilisant la définition (la formule).
   On peut procéder de deux façons, tout d'abord avec une ***approche matricielle*** *:*

```
m <- ecount(G)
D <- degree(G)
A <- as_adj(G)
D1 <- as.vector(D)
dd <- D1%*%t(D1)/(2*m)
del1 <- matrix(rep(cl1,6),ncol=6)
del2 <- matrix(rep(cl1,6),ncol=6,byrow=T)
delta <- del1==del2
sum((A-dd)*delta)/(2*m)
[1] 0.3
```

```
del1 <- matrix(rep(cl2,6),ncol=6)
del2 <- matrix(rep(cl2,6),ncol=6,byrow=T)
delta <- del1==del2
sum((A-dd)*delta)/(2*m)
[1] -0.32
```

   On peut également utiliser des ***boucles*** *(moins efficace) :*

```
ma_mod <- function(G,cl){
  m <- ecount(G)
  n <- vcount(G)
  dd <- degree(G)
  res <- 0
  A <- as_adj(G)
  for (i in 1:n){
    for (j in 1:n){
      if (cl[i]==cl[j]){
        res <- res+(A[i,j]-dd[i]*dd[j]/(2*m))
      }
```

```
    }
  }
  return(res/(2*m))
}
ma_mod(G,cl1)
[1] 0.3
ma_mod(G,cl2)
[1] -0.32
```

2. Retrouver ces deux valeurs avec la fonction **modularity**.

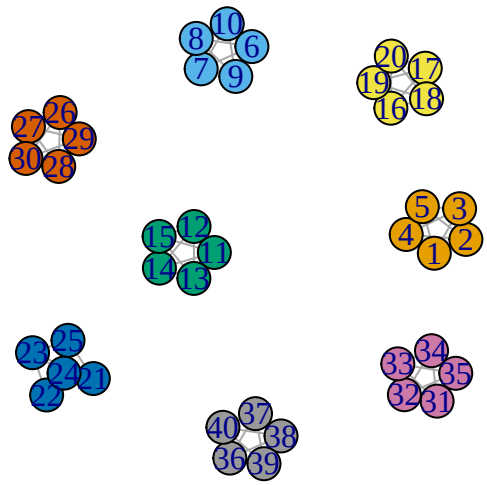
```
modularity(G,cl1)
[1] 0.3
modularity(G,cl2)
[1] -0.32
```

3. Construire un graphe et proposer une partition avec une modularité élevée et une autre avec une modularité faible.

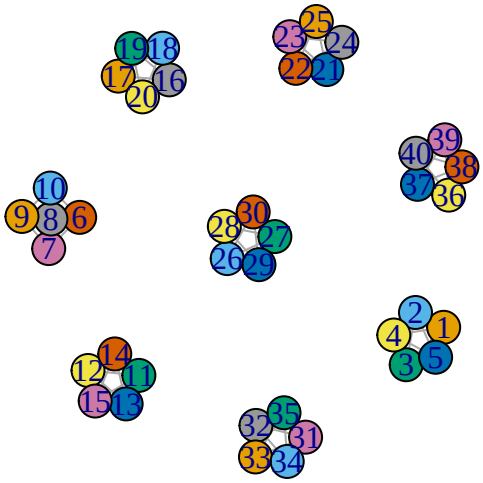
   On propose un graphe à 8 composantes connexe avec une partition "parfaite" et une autre (très) mauvaise :

```
G1 <- make_full_graph(5)+make_full_graph(5)+make_full_graph(5)+make_full_graph(5)+
  make_full_graph(5)+make_full_graph(5)+make_full_graph(5)+make_full_graph(5)
cl1 <- c(rep(1:8,each=5))
cl2 <- rep(1:8,5)
```

```
plot(G1,vertex.color=cl1)
```




```
plot(G1,vertex.color=cl2)
```

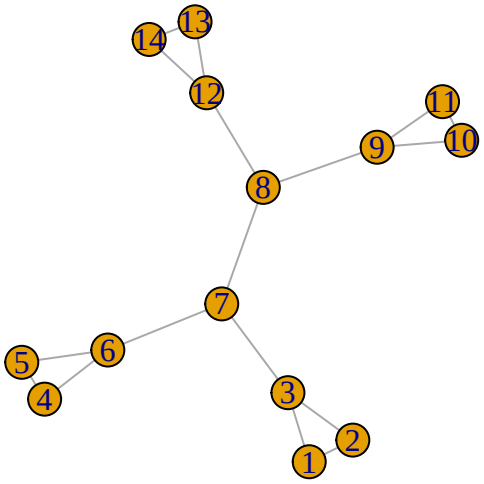



```
modularity(G1,cl1)
[1] 0.875
modularity(G1,cl2)
[1] -0.125
```

Exercice 3.2 (Edge betweeness et méthode de Louvain)**.**

On considère le graphe suivant :

```
G1 <- make_full_graph(3)
G2 <- make_full_graph(3)
G3 <- make_full_graph(2)
G4 <- make_full_graph(3)
G5 <- make_full_graph(3)
G <- G1+G2+G3+G4+G5
G <- add.edges(G, c(6,7))
G <- add.edges(G, c(3,7))
G <- add.edges(G, c(8,9))
G <- add.edges(G, c(8,12))
plot(G)
```



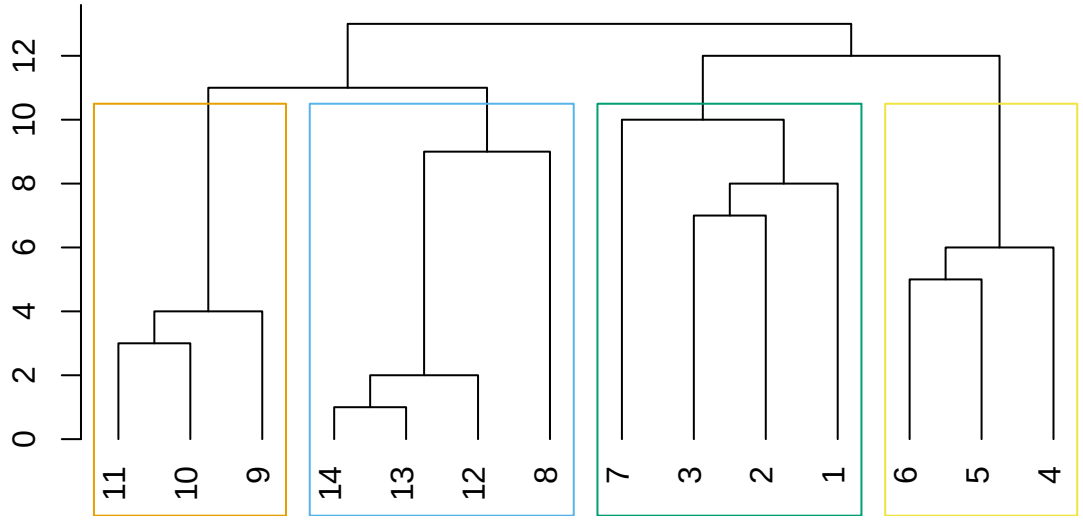
1. Calculer l'**edge betweeness** de chaque arête et identifier l'arête qui possède la plus forte valeur.

```
edge_betweenness(G)
 [1]  1 12 12  1 12 12 49 12 12  1 12 12  1 33 33 33 33
```

C'est la 7ème arête qui possède la plus forte valeur.

2. Effectuer le clustering par edge betweeness et visualiser le **dendrogramme**. Identifier la première arête retirée.

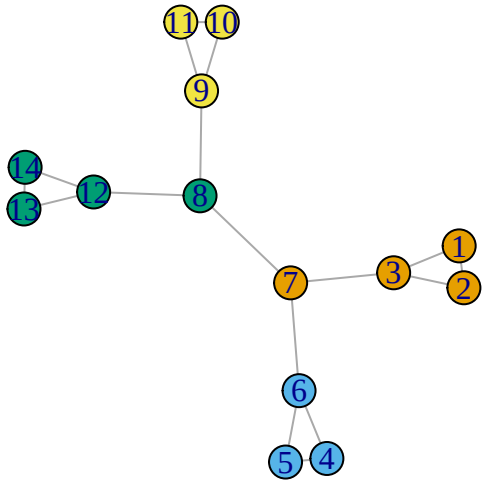
```
res <- cluster_edge_betweenness(G)
dendPlot(res)
```



C'est bien la 7ème arête qui a été retirée en premier.

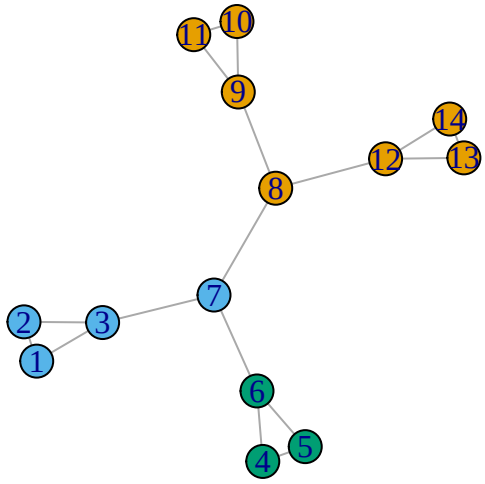
3. Représenter les classes sur le graphe.

```
plot(G,vertex.color=res$membership)
```



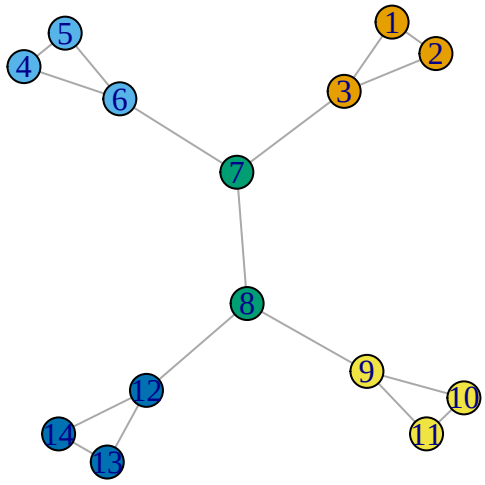
4. Couper le dendrogramme pour obtenir 3 classes. On pourra utiliser **cutat**.

```
gr3=cutat(res, no=3)
plot(G, vertex.color=gr3)
```



5. Comparer le résultat avec la méthode de Louvain.

```
set.seed(1234)
res1 <- cluster_louvain(G)
plot(G, vertex.color=res1$membership)
```



6. Comparer les modularités obtenues.

```
modularity(G,res$membership)
[1] 0.5657439
modularity(G,gr3)
[1] 0.5034602
modularity(G,res1$membership)
[1] 0.5640138
```

7. Comparer avec **cluster_optimal**.

```
#res2 <- cluster_optimal(G)
res2 <- cluster_fast_greedy(G)
modularity(G,res2$membership)
[1] 0.5640138
```

On retrouve le même résultat que l'edge betweeness qui, sur cet exemple, maximise la modularité.

Exercice 3.3 (Communautés pour karaté et friends)**.**

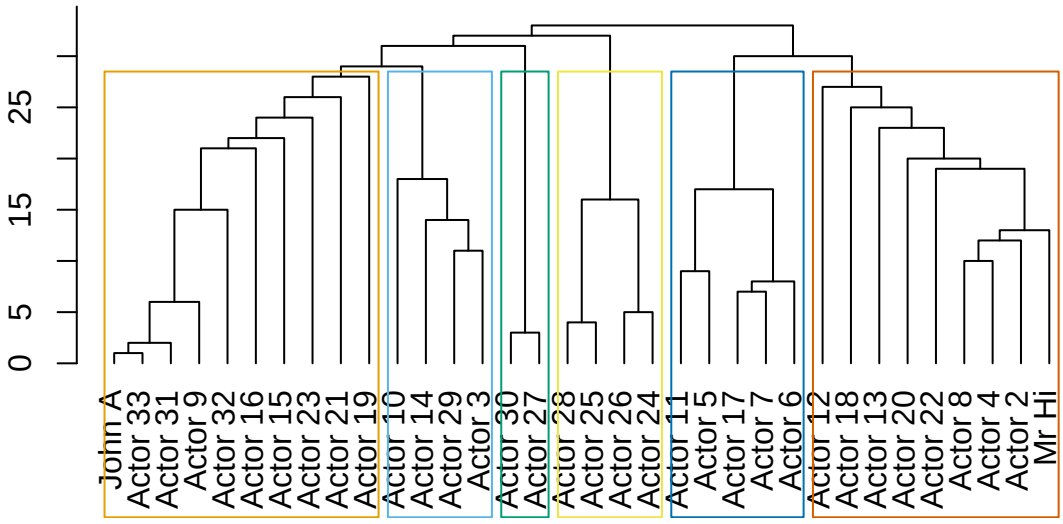
Utiliser les techniques basées sur la modularité pour faire de classes sur les données **karate** et **friends**.

On commence par le karaté :

```
library(igraphdata)
data(karate)
```

*On a pour l'**edge betweeness***

```
clust1.eb <- cluster_edge_betweenness(karate)
dendPlot(clust1.eb)
```

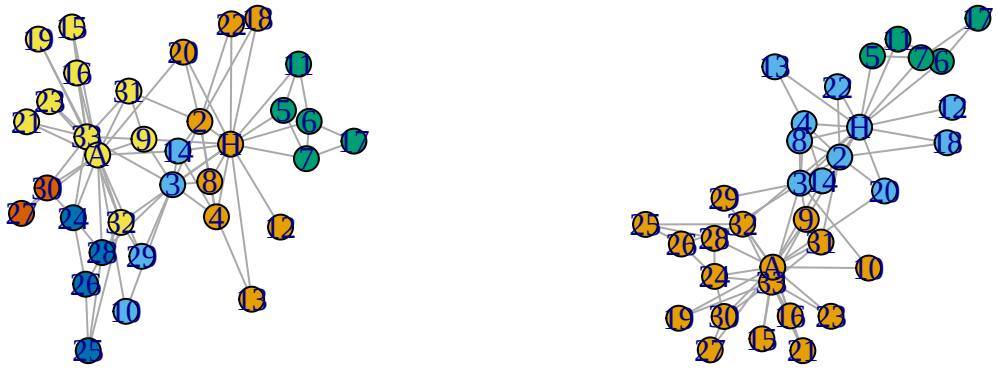


On obtient une partition en 6 classes. On maximise maintenant la modularité. Le graphe étant petit, il est possible d'être exhaustif.

```
#clust.opt <- cluster_optimal(karate)
clust.opt <- cluster_fast_greedy(karate)
```

On remarquera qu'il n'est pas possible d'afficher d'afficher le dendrogramme puisque la méthode n'est pas hiérarchique ! On peut néanmoins comparer les deux partitions :

```
par(mfrow=c(1,2))
plot(karate,vertex.color=clust1.eb$membership)
plot(karate,vertex.color=clust.opt$membership)
```



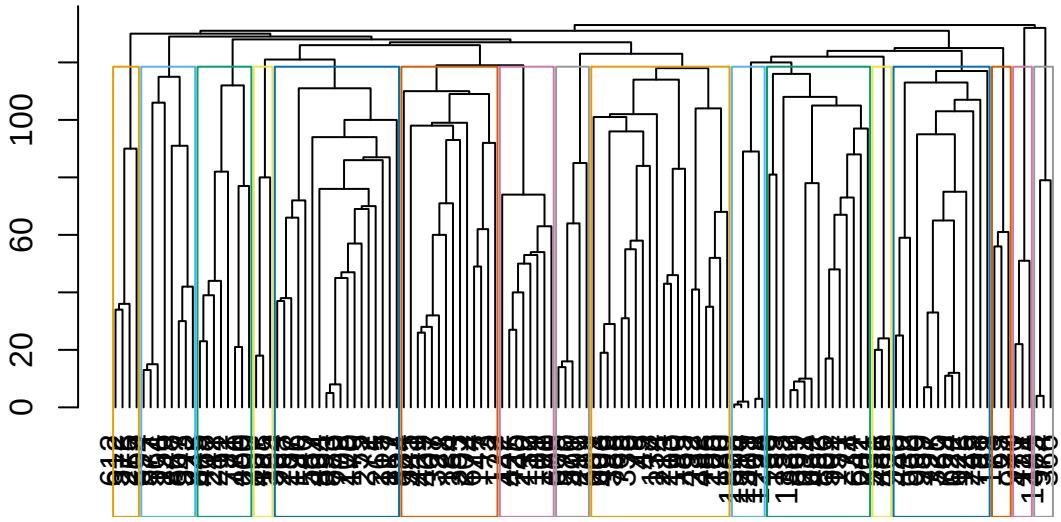
et calculer la modularité

```
modularity(karate,clust1.eb$membership)
[1] 0.3618508
modularity(karate,clust.opt$membership)
[1] 0.3990796
```

On fait le même travail pour l'autre graphe :

```
friends <- read.table(file='data/Friendship-network_data_2013.csv')
amis <- graph_from_data_frame(friends,directed=F)
```

```
clust1.eb <- cluster_edge_betweenness(amis)
dendPlot(clust1.eb)
```



Le nombre de nœuds étant assez important, le clustering optimal prend trop de temps :

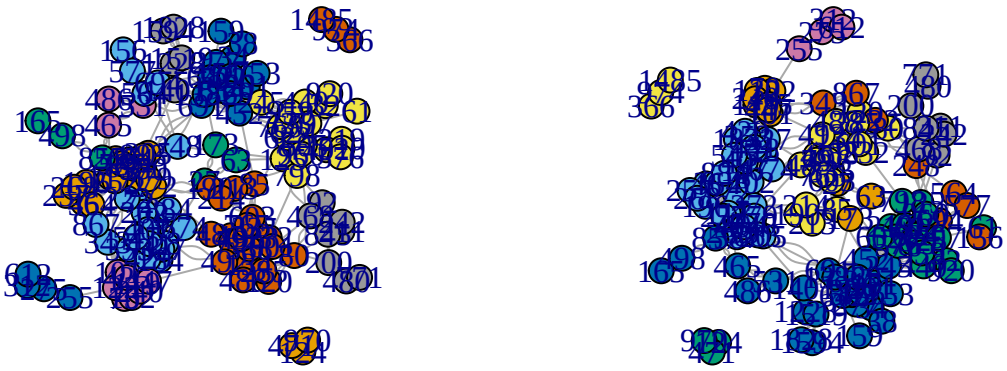
```
clust.opt <- cluster_optimal(amis)
```

On va donc plutôt utiliser la méthode de Louvain

```
clust.louv <- cluster_louvain(amis)
```

On compare les deux partitions :

```
par(mfrow=c(1,2))
plot(amis,vertex.color=clust1.eb$membership)
plot(amis,vertex.color=clust.louv$membership)
```



```
modularity(amis,clust1.eb$membership)
[1] 0.7409451
modularity(amis,clust.louv$membership)
[1] 0.7468334
```

4 Clustering spectral

Le *clustering spectral* est un algorithme de classification non supervisé qui permet de définir des clusters de nœuds sur des graphes ou d'individus pour des données `individus/variables`. L'algorithme est basé sur la décomposition spectrale du Laplacien (normalisé) d'une matrice de similarité, il est résumé ci-dessous :

Entrées :

— tableau de données $n \times p$
— K un noyau
— k le nombre de clusters.

1. Calculer la matrice de **similarités** W sur les données en utilisant le **noyau** K
2. Calculer le **Laplacien normalisé** L_{norm} à partir de W.
3. Calculer les k **premiers vecteurs propres** $u_1, \ldots, u_k$ de L_{norm}. On note U la matrice $n \times k$ qui les contient.
4. Calculer la matrice T en **normalisant les lignes** de U : $t_{ij} = u_{ij}/(\sum_\ell u_{i\ell}^2)^{1/2}$.
5. Faire un k**-means** avec les points $y_i, i = 1, \ldots, n$ (i-ème ligne de T) $\Longrightarrow A_1, \ldots, A_k$.

Sortie : clusters $C_1, \ldots, C_k$ avec

$$C_j = \{i | y_i \in A_j\}.$$

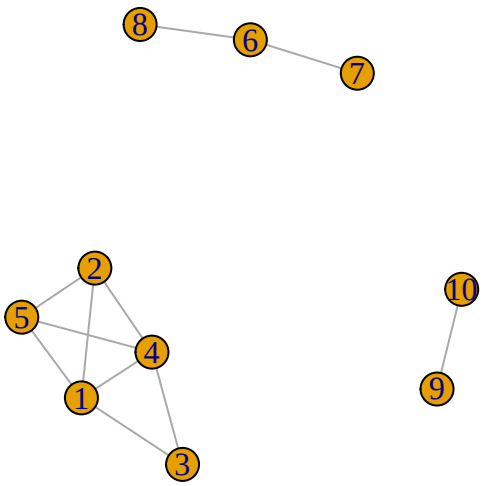
L'objet de ce chapitre est de travailler sur cet algorithme en le programmant, puis en utilisant la fonction **specc** du package `kernlab`.

4.1 Clustering spectral sur 1 graphe à 3 composantes connexes

On crée tout d'abord un graphe avec trois composantes connexes : on utilise la commande **sample_gnp()** qui permet de créer un graphe selon le modèle d'Erdos-Renyi.

```
set.seed(1)
n1 <- 5
n2 <- 3
n3 <- 2
n <- n1+n2+n3
```

```
# il faut prendre des grandes valeurs de p sinon on risque d'avoir des sous-graphes non connexes
p1 <- 0.85
p2 <- 0.75
p3 <- 0.7
G1 <- sample_gnp(n1,p1)
G2 <- sample_gnp(n2,p2)
G3 <- sample_gnp(n3,p3)
G <- G1 + G2 + G3 # il cree un graphe avec ces 3 sous-graphes
plot(G)
```



On vérifie le nombre de composantes connexes

```
components(G)$no
[1] 3
```

Exercice 4.1 (Laplacien non normalisé)**.**

1. Calculer la matrice d'adjacence de **G** et en déduire le Laplacien non normalisé.

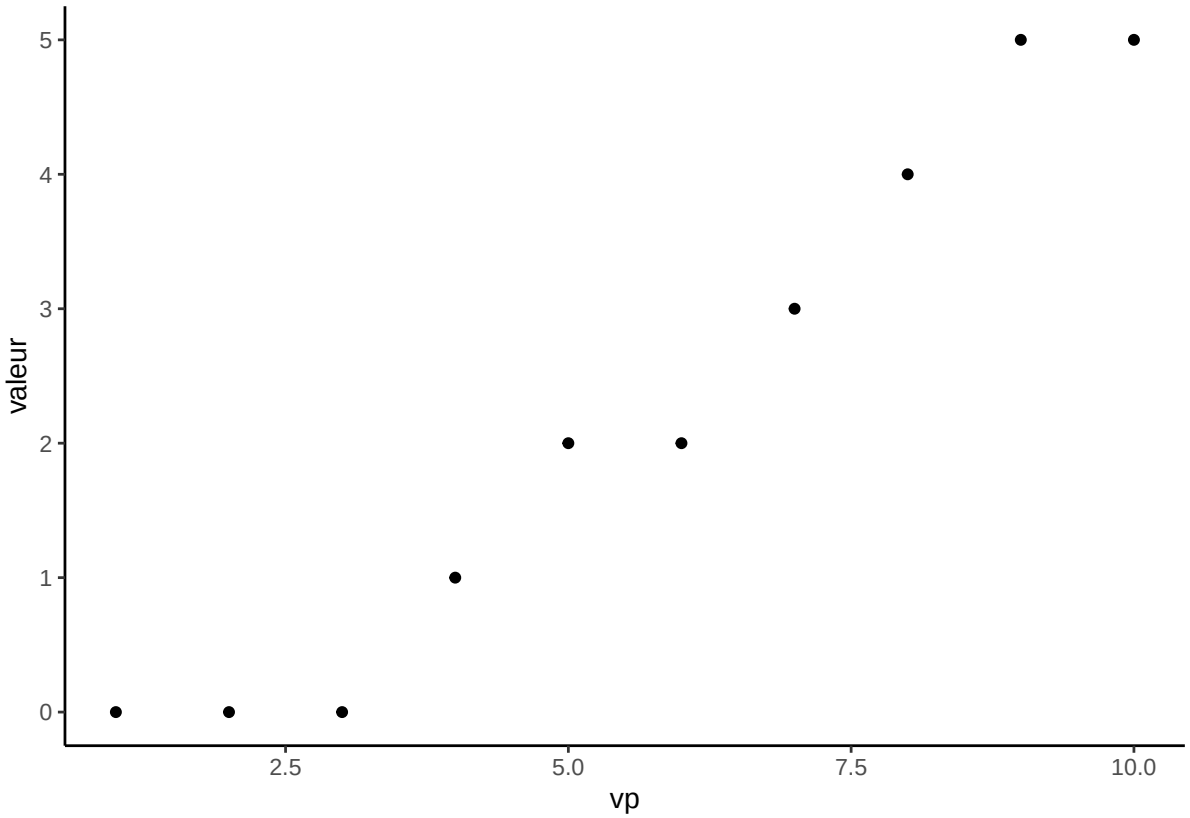
```
A <- as_adj(G,sparse=F)
D <- diag(rowSums(A))
L <- D - A
```

2. Retrouver ce Laplacien avec la fonction **laplacian_matrix**.

```
L <- laplacian_matrix(G,sparse=F)
```

3. Calculer les valeurs propres et représenter les sur un graphe. Que remarquez-vous ?

```
spec <- eigen(L)
spec$values
 [1] 5.000000e+00 5.000000e+00 4.000000e+00 3.000000e+00
 [5] 2.000000e+00 2.000000e+00 1.000000e+00 3.552714e-15
 [9] 1.776357e-15 1.110223e-15
df <- tibble(vp=1:length(spec$values),valeur=rev(spec$values))
ggplot(df)+aes(x=vp,y=valeur)+geom_point()
```



On observe bien le ***"trou spectral"*** *(eigengap) entre la troisième et la quatrième valeur propre. Conformément à la théorie, l'ordre de multiplicité de la valeur propre 0 est égal au nombre de composantes connexes du graphe.*

4. Obtenir les trois vecteurs propres associés à la valeur propre nulle. Commenter.

```
U <- spec$vectors[,n:(n-2)]
U
            [,1]       [,2]       [,3]
 [1,] 0.0000000  0.0000000 -0.4472136
 [2,] 0.0000000  0.0000000 -0.4472136
 [3,] 0.0000000  0.0000000 -0.4472136
 [4,] 0.0000000  0.0000000 -0.4472136
 [5,] 0.0000000  0.0000000 -0.4472136
 [6,] 0.0000000 -0.5773503  0.0000000
 [7,] 0.0000000 -0.5773503  0.0000000
 [8,] 0.0000000 -0.5773503  0.0000000
 [9,] 0.7071068  0.0000000  0.0000000
[10,] 0.7071068  0.0000000  0.0000000
```

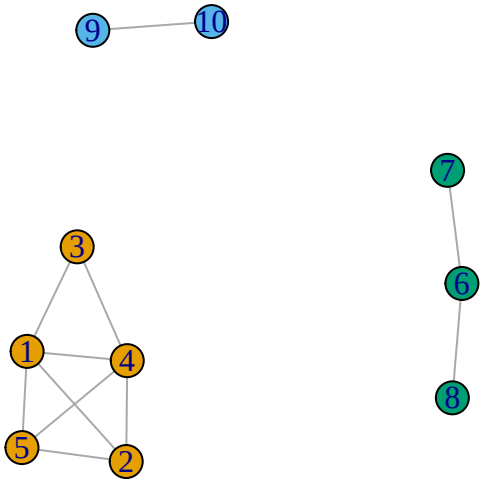
On voit que la matrice U des trois vecteurs propres (en colonne) associés à 0 est une matrice qui n'a que trois lignes différentes.

5. Terminer l'algorithme de clustering spectral avec l'étape de k means.
Le clustering de ces lignes est donc immédiat ici.

```
res <- kmeans(U,3,nstart=100)
res$cluster
 [1] 1 1 1 1 1 3 3 3 2 2
```

6. Visualiser les clusters.

```
plot(G, vertex.color=res$cluster)
```



Exercice 4.2 (Laplacien normalisé)**.**

Refaire le même travail en utilisant le laplacien normalisé. On n'oubliera pas d'ajouter l'étape de normalisation en utilisant par exemple la fonction suivante :

```
normalize <- function(x){
  return(x/sqrt(sum(x^2)))
}
```

On calcule tout d'abord le Laplacien normalisé

```
D_moins1_2 <- diag(1/sqrt(diag(D)))
LN <- diag(n) - D_moins1_2 %*% A %*%D_moins1_2
```

ou avec

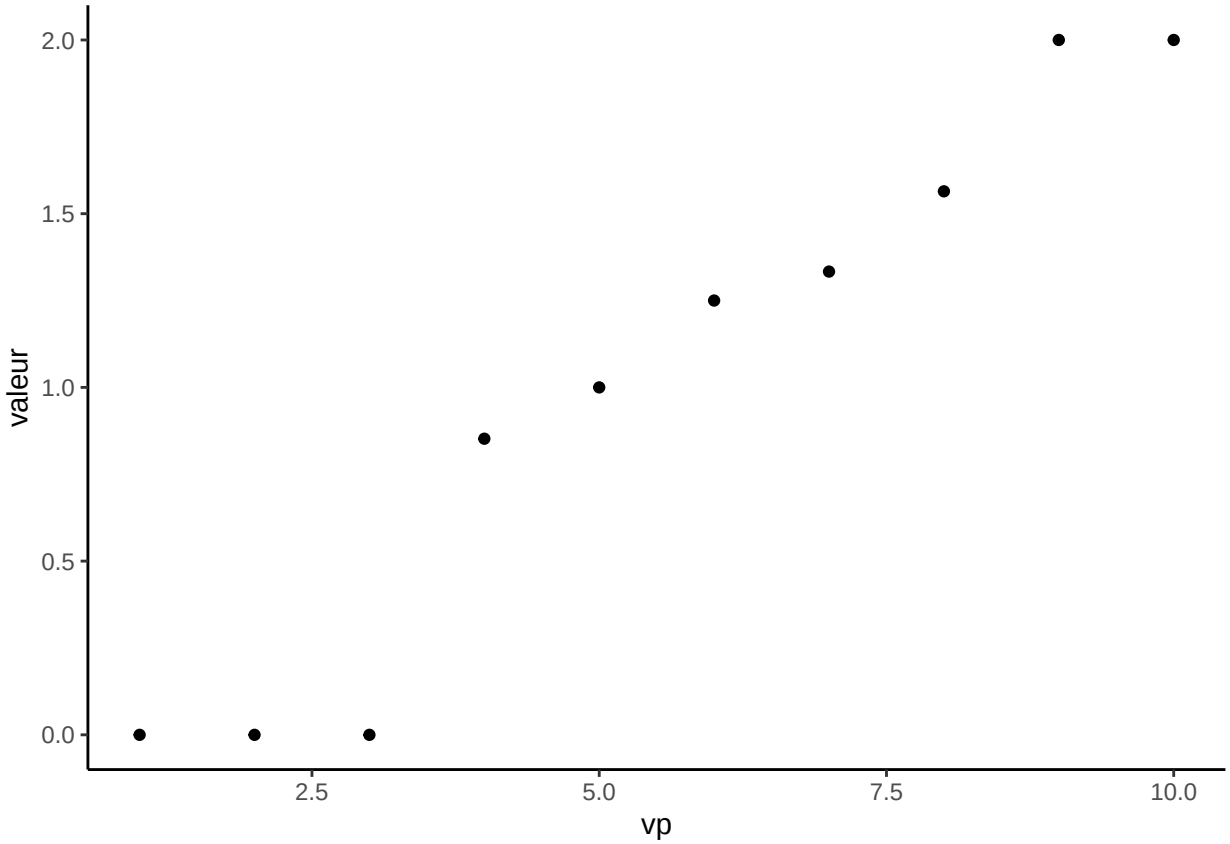
```
LN <- laplacian_matrix(G,norm=TRUE,sparse=F)
```

On calcule les valeurs propres et vecteurs propres

```
specN <- eigen(LN)
specN$values
 [1] 2.000000e+00 2.000000e+00 1.564333e+00 1.333333e+00
 [5] 1.250000e+00 1.000000e+00 8.523332e-01 1.332268e-15
 [9] 1.110223e-15 1.110223e-15
```

et on observe à nouveau un trou spectral entre les valeurs propres 3 et 4.

```
dfN <- tibble(vp=1:length(specN$values),valeur=rev(specN$values))
ggplot(dfN)+aes(x=vp,y=valeur)+geom_point()+theme_classic()
```



On calcule les 3 vecteurs propres :

```
U <- specN$vectors[,n:(n-2)]
U
            [,1]      [,2]       [,3]
 [1,] 0.0000000 0.0000000 -0.5000000
 [2,] 0.0000000 0.0000000 -0.4330127
 [3,] 0.0000000 0.0000000 -0.3535534
 [4,] 0.0000000 0.0000000 -0.5000000
 [5,] 0.0000000 0.0000000 -0.4330127
 [6,] 0.7071068 0.0000000  0.0000000
 [7,] 0.5000000 0.0000000  0.0000000
 [8,] 0.5000000 0.0000000  0.0000000
 [9,] 0.0000000 0.7071068  0.0000000
[10,] 0.0000000 0.7071068  0.0000000
```

que l'on normalise à l'aide de la fonction ***normalize*** *:*

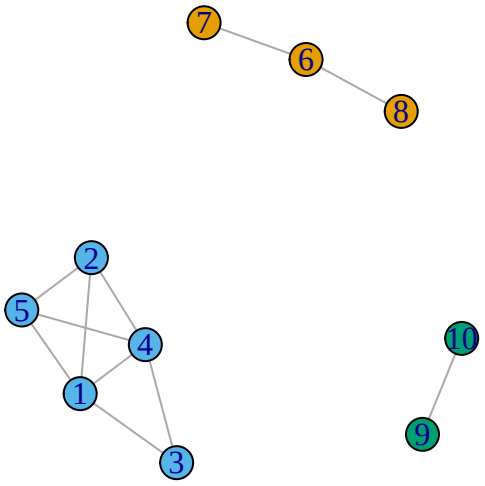
```
U.norm <- t(apply(U,1,normalize))
U.norm
      [,1] [,2] [,3]
 [1,]    0    0   -1
```

```
 [2,]    0    0   -1
 [3,]    0    0   -1
 [4,]    0    0   -1
 [5,]    0    0   -1
 [6,]    1    0    0
 [7,]    1    0    0
 [8,]    1    0    0
 [9,]    0    1    0
[10,]    0    1    0
```

Il reste à faire le k-means.

```
res <- kmeans(U.norm,3,nstart=100)
res$cluster
 [1] 2 2 2 2 2 1 1 1 3 3
```

```
plot(G, vertex.color=res$cluster)
```



4.2 Programmer le clustering spectral pour un graphe

Exercice 4.3 (Construction de l'algorithme)**.**

Créer une fonction **R** qui admet en entrée :

— un graphe
— une valeur de K (un entier positif)

et qui renvoie les groupes pour le clustering spectral à K groupes ainsi que le graphe des valeurs propres (en **ggplot** si possible).

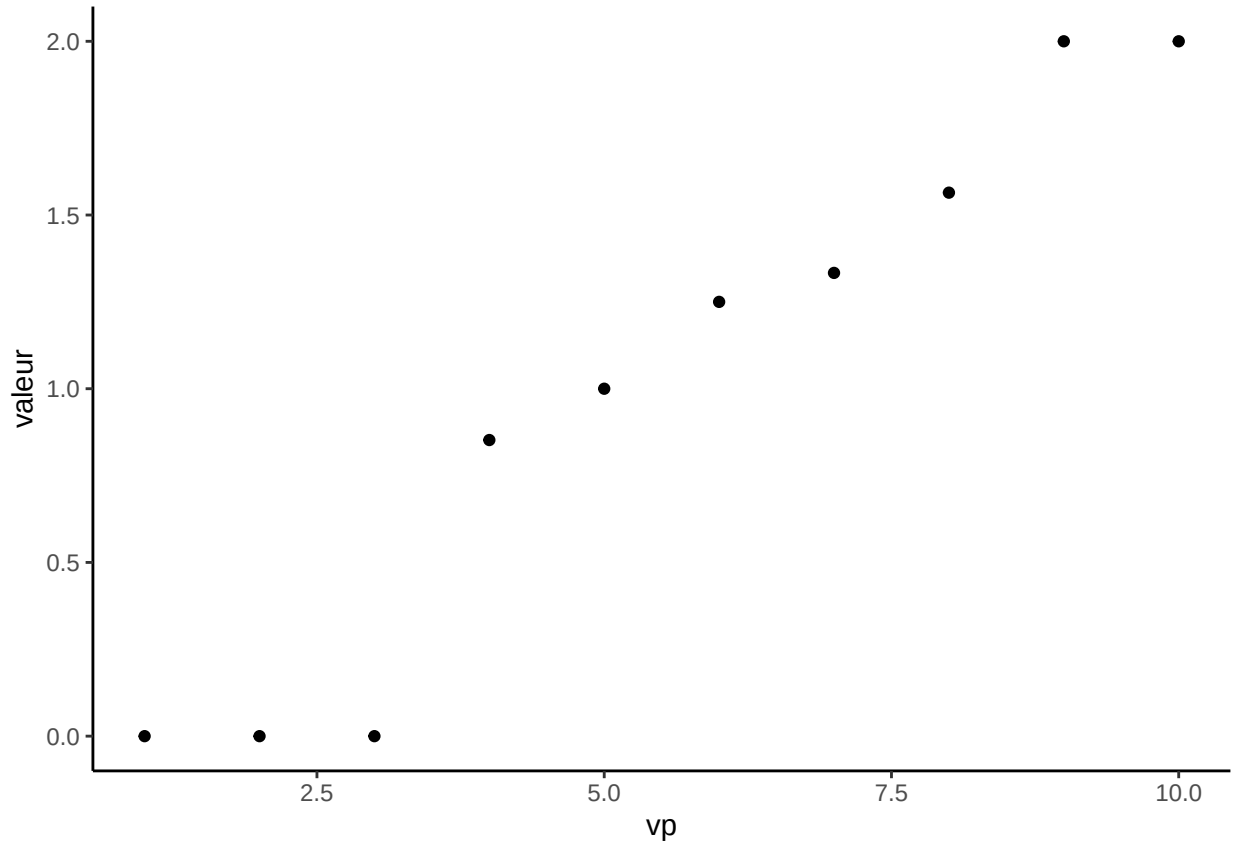
Il suffit de mettre en forme ce qui a été fait dans la partie précédente, par exemple :

```
spec.clust.norm <- function(G,K){
  n <- vcount(G)
  LN <- laplacian_matrix(G,norm=TRUE,sparse=F)
  specN <- eigen(LN)
  dfN <- tibble(vp=1:length(specN$values),valeur=rev(specN$values))
```

```
  graph <- ggplot(dfN)+aes(x=vp,y=valeur)+geom_point()+theme_classic()
  U <- specN$vectors[,n:(n-K+1)]
  U.norm <- t(apply(U,1,normalize))
  clustering <- kmeans(U.norm,K,nstart=100)$cluster
  return(list(groupe=clustering,graphe=graph,valeur=rev(specN$value)))
}
```

On teste la fonction :

```
res <- spec.clust.norm(G,K=8)
res$graphe
```




```
res1 <- spec.clust.norm(G,K=3)
res1$groupe
 [1] 2 2 2 2 2 3 3 3 1 1
```

Igraph possède une fonction permettant de faire directement le spectral clustering : **embed_laplacian_matrix**. Mais en argument, il faut lui donner le nombre K de clusters souhaité. En pratique, on ne connait pas K, et une façon de le trouver est de regarder le trou spectral dans le graphe des valeurs propres.

La fonction **embed_laplacian_matrix** s'utilise ainsi :

```
res2 <- embed_laplacian_matrix(G,8,which="sa",scaled="FALSE",degmode = "all")
res2$D
[1] 0 0 0 1 2 2 3 4
```

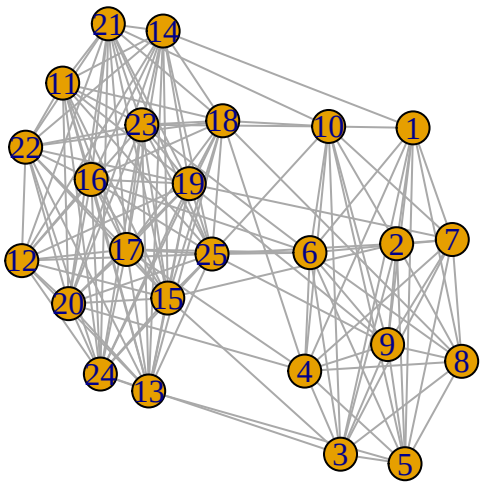
On peut faire du clustering spectral à 3 groupes avec :

```
res3 <- embed_laplacian_matrix(G,3,which="sa",scaled="FALSE",degmode = "all")
res_spectral <- kmeans(res3$X, centers = 3, nstart = 100)
res_spectral$cluster
 [1] 2 2 2 2 2 3 3 3 1 1
```

Exercice 4.4 (Graphe avec deux communautés faiblement connectées entre elles)**.**

On considère le graphe suivant obtenu selon un modèle **SBM** :

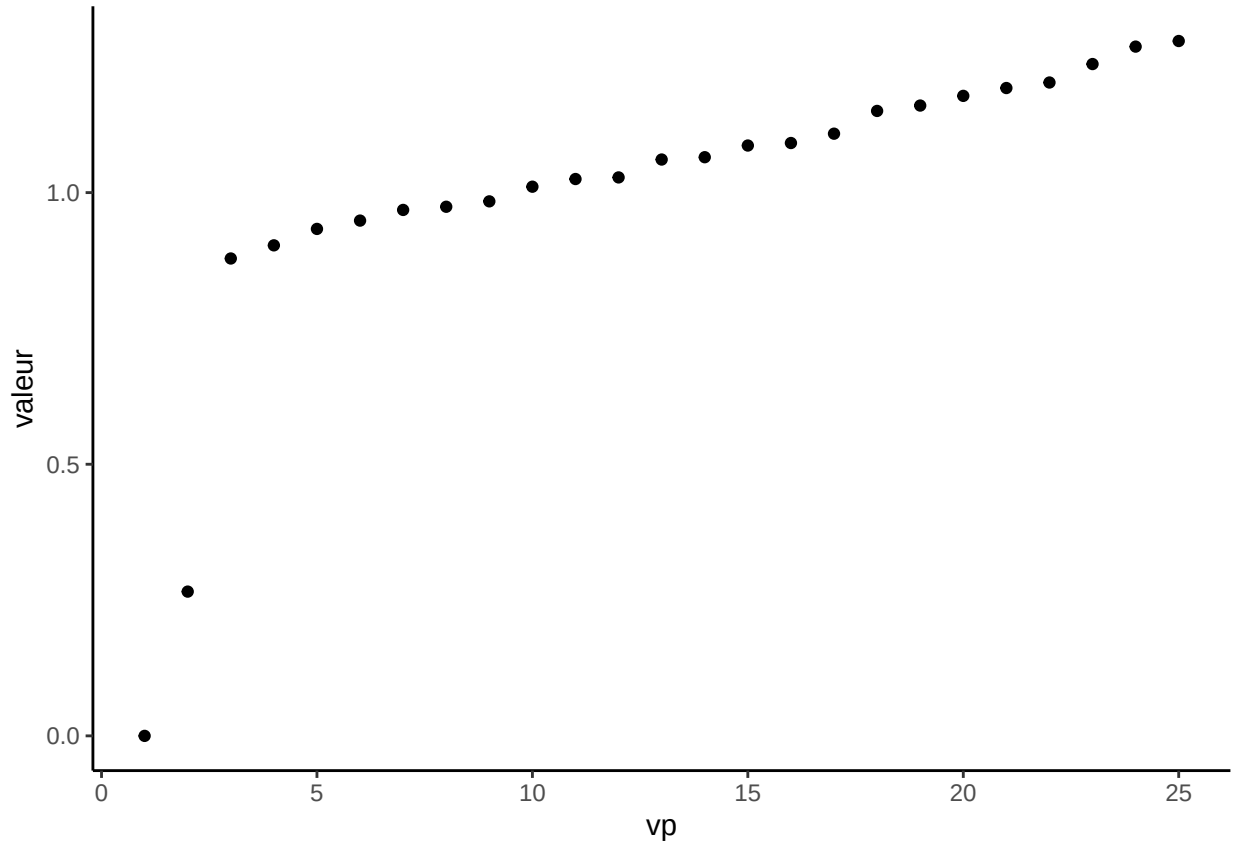
```
set.seed(1234)
n <- 25# nombre de noeuds
Q <- 2 # nombre de cluster clusters
pi <- c(0.4, 0.6) # taille des groupes
effectifs <- n*pi
connectivite_matrix <- matrix(c(0.9, 0.15,
                                0.15, 0.95),nrow=Q) # matrice de connexion
G <- sample_sbm(n, pref.matrix=connectivite_matrix, block.sizes = effectifs)
plot(G)
```



Effectuer le clustering spectral sur ce graphe, on essaiera notamment de choisir le nombre de groupes.

On commence avec un nombre de groupes "grand" pour visualiser le graphe des valeurs propres :

```
res <- spec.clust.norm(G,K=8)
res$graphe
```

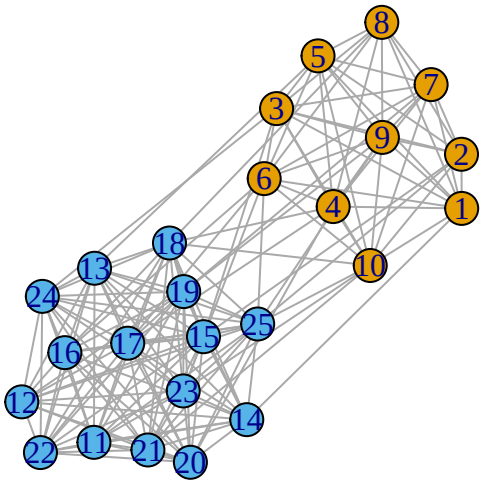


Il n'y a plus qu'une valeur propre nulle, c'est normal puisqu'on a une seule composante connexe. On remarque néanmoins un ***trou spectral*** *après la seconde valeur propre, ce qui laisse supposer un graphe à 2 groupes. On refait donc le clustering spectral avec* $K = 2$

```
res <- spec.clust.norm(G,K=2)
groupe <- res$groupe
```

On visualise les groupes :

```
plot(G,vertex.color=groupe)
```



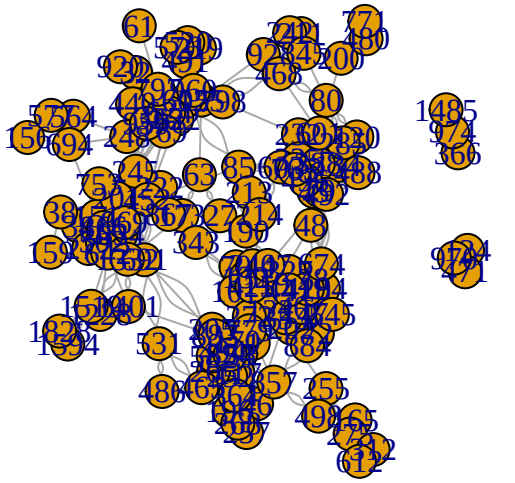
L'algorithme a bien identifié les deux groupes.

4.3 Exemple sur des graphes "réels"

Exercice 4.5 (Clustering spectral sur deux graphes)**.**

1. On considère le graphe **friends** disponible ici

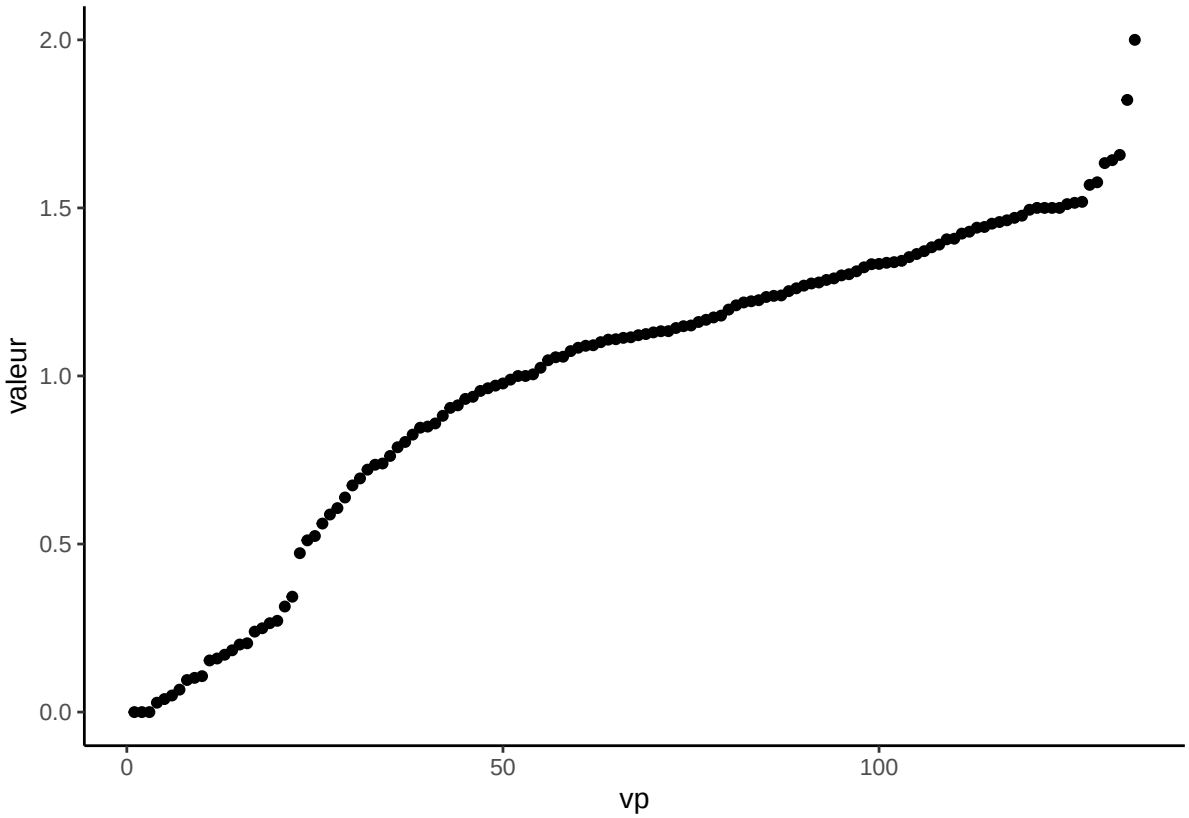
```
friends <- read.table(file="data/Friendship-network_data_2013.csv")
G.friends <- graph_from_data_frame(friends,directed=F)  # non dirige
plot(G.friends)
```



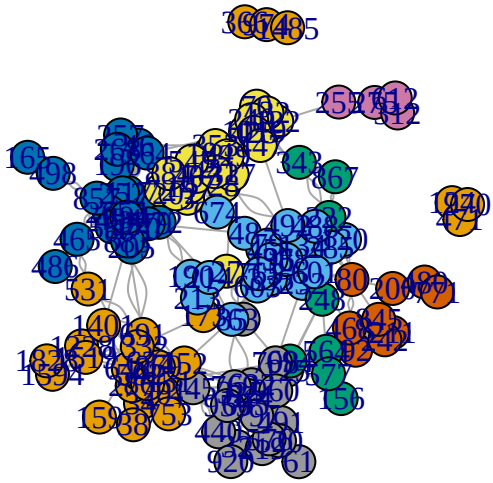
Appliquer le clustering spectral à ce graphe. On pourra comparer la classification obtenue avec celle de la méthode de Louvain en utilisant la fonction **compare**.

On applique l'algorithme avec 8 groupes

```
res1 <- spec.clust.norm(G.friends,K=8)
res1$graphe
```

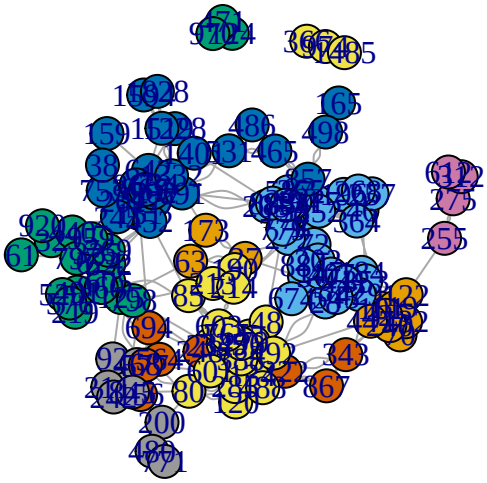



```
res2 <- spec.clust.norm(G.friends,K=9)
plot(G.friends,vertex.color=res2$groupe)
```



On maximise maintenant la ***modularité*** *avec la méthode de* ***Louvain****.*

```
set.seed(1234)
cl.mod <- cluster_louvain(G.friends)
plot(G.friends,vertex.color=cl.mod$membership)
```



On obtient 13 communautés

```
max(cl.mod$membership)
[1] 13
```

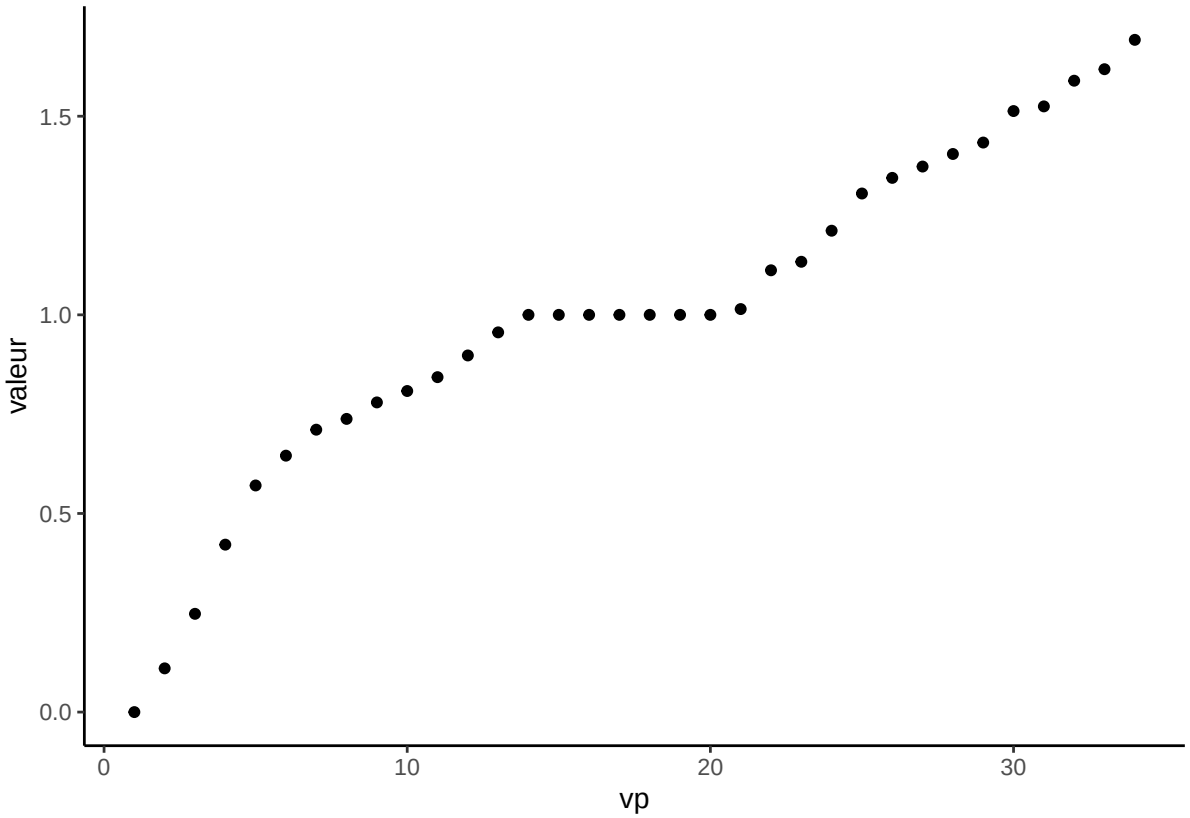
On compare ces deux classifications avec ***l'indice de Rand ajusté****, voir par exemple ici*

```
compare(res2$groupe,cl.mod$membership,method="adjusted.rand")
[1] 0.789699
```

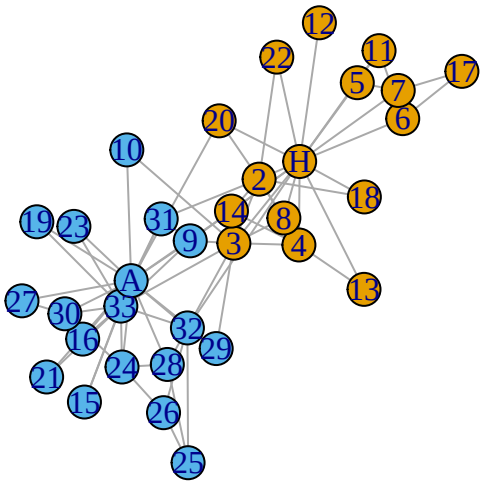
2. Faire de même avec le graphe **karate**.

```
library(igraphdata)
data(karate)
```

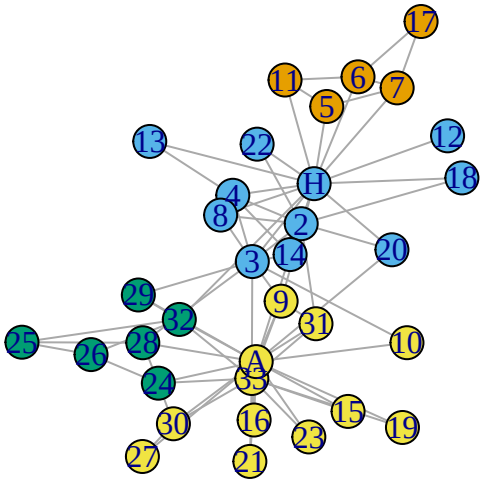
```
res1 <- spec.clust.norm(karate,K=8)
res1$graphe
```




```
res2 <- spec.clust.norm(karate,K=2)
plot(karate,vertex.color=res2$groupe)
```




```
cl.louv <- cluster_louvain(karate)
plot(karate,vertex.color=cl.louv$membership)
```



4.4 Clustering spectral : cas général

Nous avons étudié jusqu'ici l'algorithme du clustering spectral pour trouver des clusters de nœuds (ou communautés) dans les graphes. On remarque néanmoins que l'algorithme ne repose pas sur le graphe en lui même, mais uniquement sur une matrice d'adjacence (ou similarité) issue de ce graphe. Il est par conséquent possible d'utiliser cet algorithme pour des données standards (tableaux `individus-variables`), à partir du moment où on peut calculer une matrice de similarité à partir de ces données. Il est également possible d'utiliser des **noyaux** pour définir cette similarité. La fonction **specc** de `kernlab` permet de faire un tel clustering.

Exercice 4.6 (Clustering spectral pour des psirales)**.**

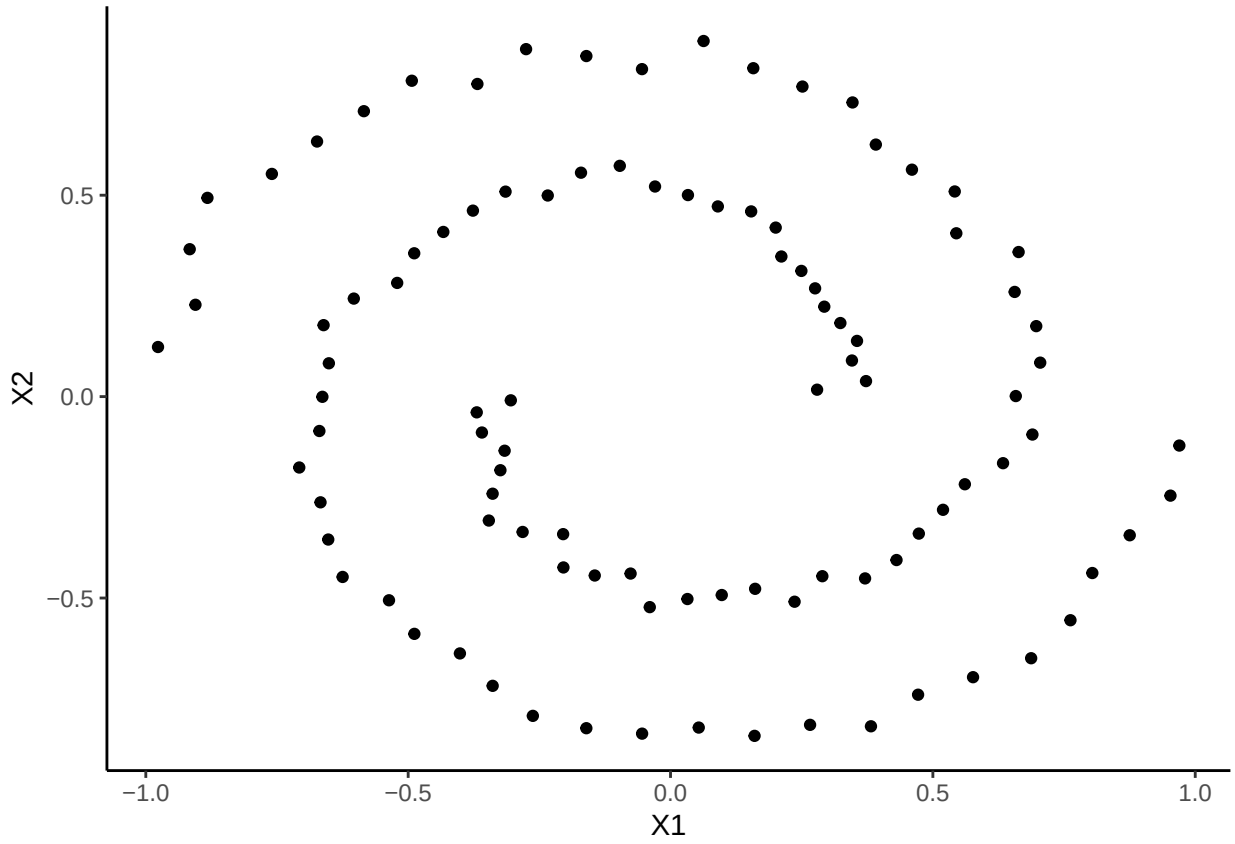
On considère les données spirales

```
set.seed(111)
library(mlbench)
n <- 100
simu <- mlbench.spirals(100,1,0.025)
names(simu)
[1] "x"       "classes"
```

```
data <- simu$x
head(data)
           [,1]        [,2]
[1,]  0.5609898 -0.21756239
[2,]  0.2793522  0.01718273
[3,]  0.3725821  0.03849122
[4,]  0.3457879  0.08963081
[5,]  0.1577921  0.81528541
[6,] -0.1603697  0.84547763
```

et on les visualise.

```
df <- data.frame(simu$x)
ggplot(df)+aes(x=X1,y=X2)+geom_point()
```



Appliquer les algorithmes suivants pour tenter de visualiser les deux groupes :

- clustering spectral avec noyau linéaire
- clustering spectral avec noyau polynomial de degree 2
- clustering spectral avec noyau radial
- k-means
- CAH avec single linkage
- CAH avec lien de Ward

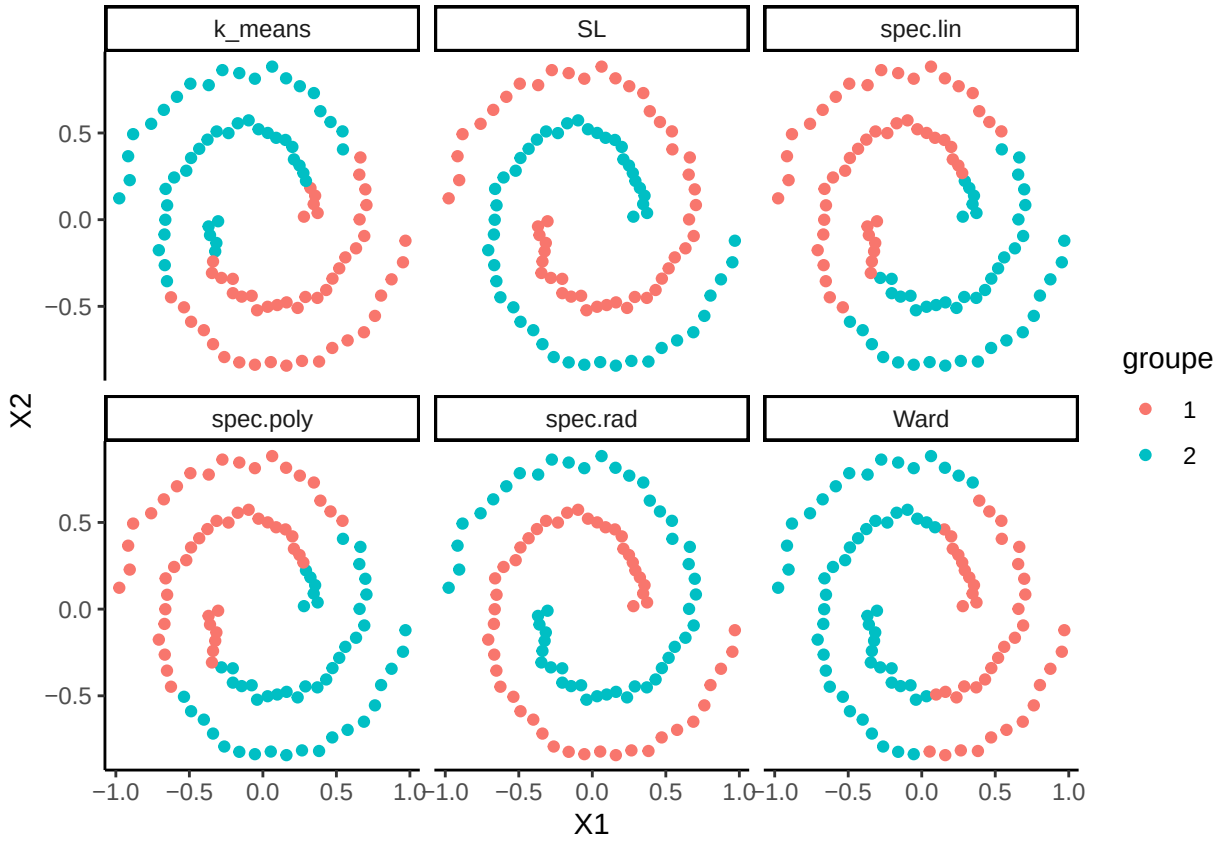
On effectue les classification avec 2 groupes :

```
set.seed(1234)
library(kernlab)
spec.lin <- specc(data,centers=2,kernel="polydot",kpar=list(degree=1))
spec.poly <- specc(data,centers=2,kernel="polydot",kpar=list(degree=2))
spec.rad <- specc(data,centers=2,kernel="rbfdot")
k_means <- kmeans(data,centers=2)
cah.ward <- hclust(dist(data),method="ward.D") %>% cutree(2) %>% as.factor()
cah.SL <- hclust(dist(data),method="single") %>% cutree(2) %>% as.factor()
```

Il reste à assembler les résultats et à les visualiser :

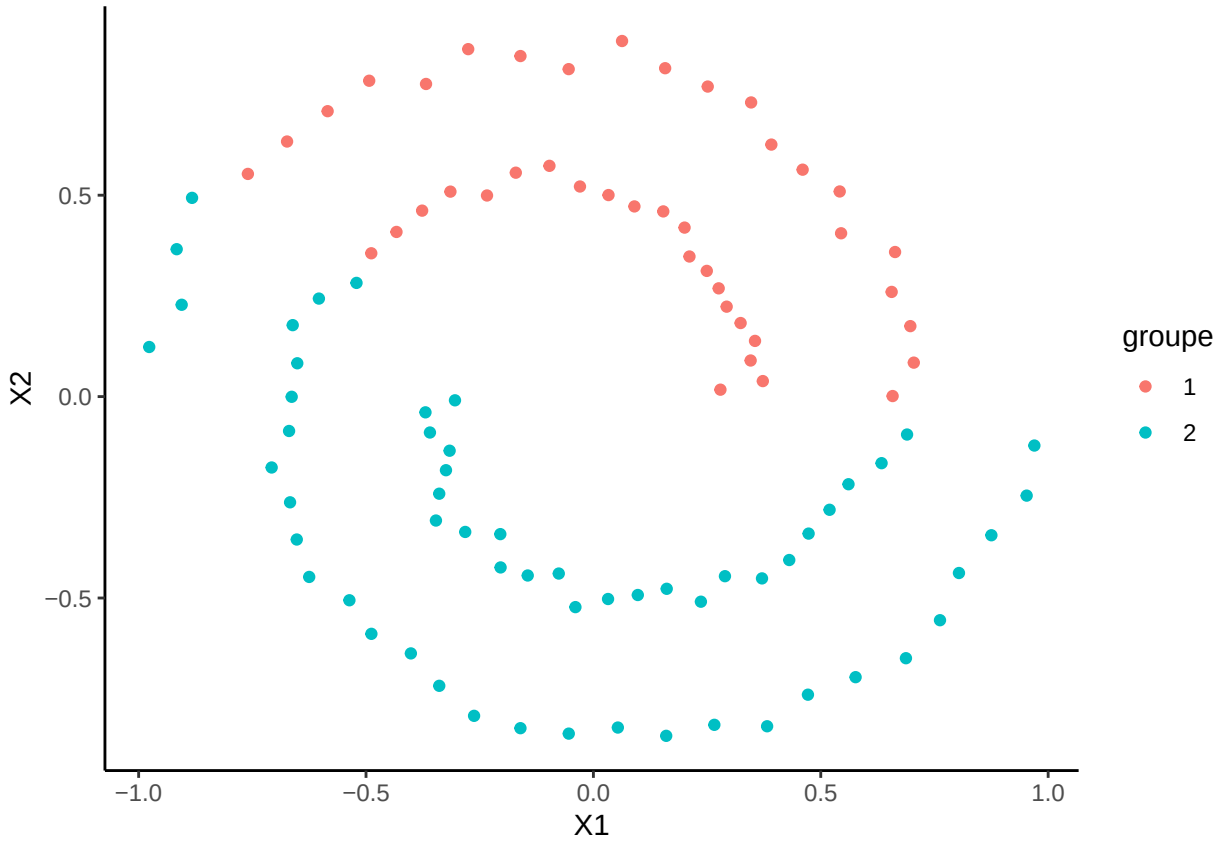
```
df1 <- df %>%
  mutate(spec.lin=as.factor(spec.lin),spec.poly=as.factor(spec.poly),
         spec.rad=as.factor(spec.rad),k_means=as.factor(k_means$cluster),
         Ward=cah.ward,SL=cah.SL) %>%
  pivot_longer(-c(X1,X2),names_to="Methode",values_to="groupe")
```

```
ggplot(df1)+aes(x=X1,y=X2,color=groupe)+geom_point()+facet_wrap(~Methode)
```



On remarque que seuls le ***single linkage*** *et le* ***clutering spectral*** *à noyau radial parviennent à identifier les deux spirales. Les autres algorithmes sont mis en échec. Pour terminer, on précise qu'il est également possible d'utiliser l'astuce du noyau avec l'algorithme du* ***kmeans****, on peut par exemple utiliser la fonction* ***kkmeans*** *de* ***kernlab*** *:*

```
km <- kkmeans(data,center=2,kernel="rbfdot",kpar="automatic")
Using automatic sigma estimation (sigest) for RBF or laplace kernel
df %>% mutate(groupe=as.factor(km)) %>% ggplot()+aes(x=X1,y=X2,color=groupe)+geom_point()
```



Ça n'apporte pas grand chose sur cet exemple.