Assignment 2 - Social Network Analysis

Part I

Start by installing the "igraph" package. Once you have installed igraph, load the package.

Now upload the data file "discipline-data.csv" as a data frame called "D1". Each row is a disciplinary action from a teacher to a student so the first line shows that teacher "E" sent student "21" to the principal. It also shows the gender of both the teacher and student and the student's main elective field of study ("major"") and the field that the teacher instructs in ("t.expertise").

```
library(igraph)
```

```
## Warning: package 'igraph' was built under R version 3.5.2
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

```
D1<-read.csv("~/Desktop/master fall/hudk4050/assignment2/discipline-data.csv")
D1 <- data.frame(D1)
```

Before you proceed, you will need to change the data type of the student id variable. Since it is a number R will automatically think it is an integer and code it as such (look at the list of variables by clicking on the data frame arrow in the Data pane. Here you will see the letters "int"" next to the stid variable, that stands for integer). However, in this case we are treating the variable as a category, there is no numeric meaning in the variable. So we need to change the format to be a category, what R calls a "factor". We can do this with the following code:

```
D1$stid <- as.factor(D1$stid)
```

igraph requires data to be in a particular structure. There are several structures that it can use but we will be using a combination of an "edge list" and a "vertex list". As you might imagine the edge list contains a list of all the relationships between students and teachers and any characteristics of those edges that we might be interested in. There are two essential variables in the edge list a "from" variable and a "to" variable that descibe the relationships between vertices (a disciplinary action is given "from" and teacher "to" a student). While the vertix list contains all the characteristics of those vertices, in our case gender and major.

So let's convert our data into an edge list!

First we will isolate the variables that are of interest: tid and stid

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:igraph':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
D2 <- select(D1, tid, stid)
```

Since our data represnts every time a teacher sends a student to the principal there are multiple rows when the same teacher sends the same student. We want to collapse these into a single row, with a variable that shows how many times a teacher-student pair appears.

```
EDGE <- count(D2, tid, stid)

names(EDGE) <- c("from", "to", "count")
EDGE
```

```
## # A tibble: 41 x 3
##    from  to    count
##    <fct> <fct> <int>
##  1 A     2         1
##  2 A     3         1
##  3 A     4         2
##  4 A     5         1
##  5 A     18        1
##  6 A     22        1
##  7 A     25        2
##  8 B     1         2
##  9 B     5         1
## 10 B     8         1
## # ... with 31 more rows
```

EDGE is your edge list. Now we need to make the vertex list, a list of all the teachers and students and their characteristics in our network.

```
#First we will separate the teachers from our original data frame
V.TCH <- select(D1, tid, t.gender, t.expertise)
#Remove all the repeats so that we just have a list of each teacher and their characteristics
V.TCH <- unique(V.TCH)
#Add a variable that describes that they are teachers
V.TCH$group <- "teacher"

#Now repeat this process for the students
V.STD <- select(D1, stid, s.gender, s.major)
V.STD <- unique(V.STD)
V.STD$stid <- as.factor(V.STD$stid)
V.STD$group <- "student"

#Make sure that the student and teacher data frames have the same variables names
names(V.TCH) <- c("id", "gender", "topic", "group")
names(V.STD) <- c("id", "gender", "topic", "group")

#Bind the two data frames together (you will get a warning because the teacher data frame has 5 types o
VERTEX <- bind_rows(V.TCH, V.STD)
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```

```
VERTEX
```

```
##    id gender   topic   group
## 1   E female     art teacher
## 2   A female biology teacher
## 3   D female     art teacher
## 4   B   male    math teacher
## 5   C   male biology teacher
## 6  21 female     art student
## 7   4   male    math student
## 8  25   male biology student
## 9  15 female english student
## 10  1   male biology student
## 11  8   male english student
## 12 10   male     art student
## 13 26 female     art student
## 14 23 female     art student
## 15  3   male     art student
## 16 27 female     art student
## 17 29 female     art student
## 18  5   male     art student
## 19 18   male biology student
## 20 28   male    math student
## 21 11   male     art student
## 22 17   male biology student
## 23  9   male biology student
## 24 12 female     art student
## 25 24   male     art student
## 26 22 female biology student
## 27  2 female    math student
## 28 19 female biology student
## 29  6   male biology student
## 30 20   male biology student
```

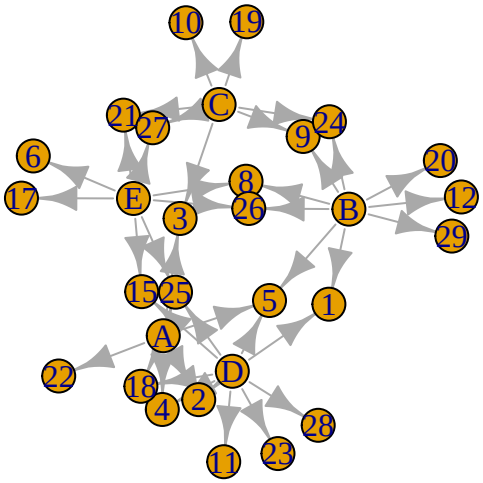
Now we have both a Vertex and Edge list it is time to plot our graph!

```
#Load the igraph package

library(igraph)

#First we will make an object that contains the graph information using our two dataframes EDGE and VER
```

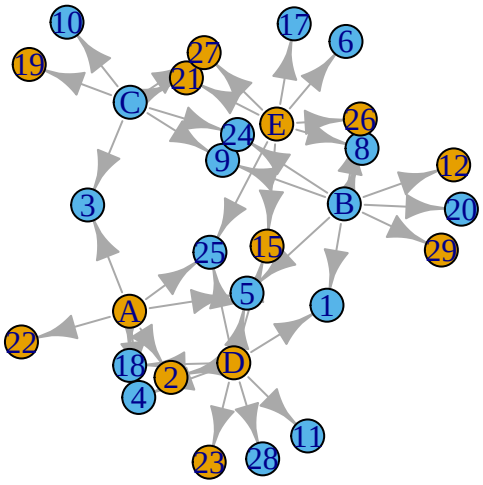
```
g <- graph.data.frame(EDGE, directed=TRUE, vertices=VERTEX)
g
```

```
## IGRAPH af146a0 DN-- 30 41 --
## + attr: name (v/c), gender (v/c), topic (v/c), group (v/c), count
## | (e/n)
## + edges from af146a0 (vertex names):
##  [1] A->2  A->3  A->4  A->5  A->18 A->22 A->25 B->1  B->5  B->8  B->9
## [12] B->12 B->20 B->24 B->26 B->29 C->3  C->9  C->10 C->19 C->21 C->24
## [23] C->27 D->1  D->2  D->4  D->5  D->11 D->15 D->18 D->23 D->25 D->28
## [34] E->6  E->8  E->15 E->17 E->21 E->25 E->26 E->27
```

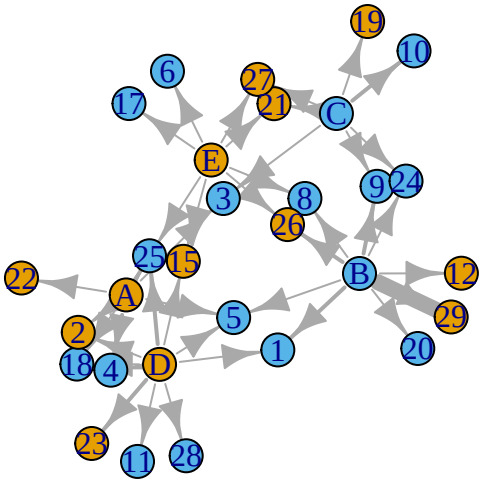
```
#Now we can plot our graph using the force directed graphing technique - our old friend Fruchertman-Rei

plot(g,layout=layout.fruchterman.reingold)
```




```
#There are many ways to change the attributes of the graph to represent different characteristics of th

plot(g,layout=layout.fruchterman.reingold, vertex.color=VERTEX$gender)
```




```
#We can change the thickness of the edge according to the number of times a particular teacher has sent

plot(g,layout=layout.fruchterman.reingold, vertex.color=VERTEX$gender, edge.width=EDGE$count)
```



Part II

In Part II your task is to look up in the igraph documentation and create a graph that sizes the student vertices in terms of the number of disciplinary actions they have recieved, and the teachers in terms of the number of disciplinary actions they have given out.

```
countstd <- EDGE %>% group_by(to)%>%summarise(sum(count))
counttch <- EDGE %>% group_by(from)%>%summarise(sum(count))
names(countstd) <- c("id","count")
names(counttch) <- c("id","count")
VERTEX1<- bind_rows(countstd,counttch)
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

```
## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```

```
## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```

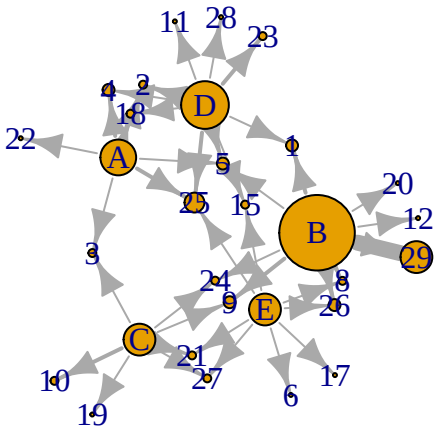
```
VERTEX1
```

```
## # A tibble: 30 x 2
##    id    count
##    <chr> <int>
##  1 1         3
##  2 2         2
##  3 3         2
##  4 4         3
##  5 5         3
##  6 6         1
##  7 8         2
##  8 9         3
##  9 10        2
## 10 11        1
## # ... with 20 more rows
```

```
g <- graph.data.frame(EDGE, directed=TRUE, vertices=VERTEX1)
g
```

```
## IGRAPH 8a9a8bf DN-- 30 41 --
## + attr: name (v/c), count (v/n), count (e/n)
## + edges from 8a9a8bf (vertex names):
##  [1] A->2  A->3  A->4  A->5  A->18 A->22 A->25 B->1  B->5  B->8  B->9
## [12] B->12 B->20 B->24 B->26 B->29 C->3  C->9  C->10 C->19 C->21 C->24
## [23] C->27 D->1  D->2  D->4  D->5  D->11 D->15 D->18 D->23 D->25 D->28
## [34] E->6  E->8  E->15 E->17 E->21 E->25 E->26 E->27
```

```
plot(g,layout=layout.fruchterman.reingold,edge.width=EDGE$count,vertex.size = VERTEX1$count*2)
```



Part III

Now practice with data from our class. Please create a **person-network** with the data set hudk4050-classes.csv. To create this network you will need to create a person-class matrix using the tidyr functions and then create a person-person matrix using `t()`. You will then need to plot a matrix rather than a data frame using igraph.

Once you have done this, also look up how to generate the following network metrics: betweeness centrality and dregree. **Who is the most central person in the network?**

```
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:igraph':
##
##     crossing
```

```
#modify data to get revelent variables
D3 <- read.csv("~/Desktop/master fall/hudk4050/assignment2/hudk4050.classes.csv")
D3 <- data.frame(D3)
names(D3) <- c("FirstName", "LastName", "Class1", "Class2","Class3","Class4","Class5","Class6")
D4 <- select(D3,FirstName,Class1,Class2,Class3,Class4,Class5,Class6)
D6<-  gather(D4,coursenum,course,Class1,Class2,Class3,Class4,Class5,Class6)
```

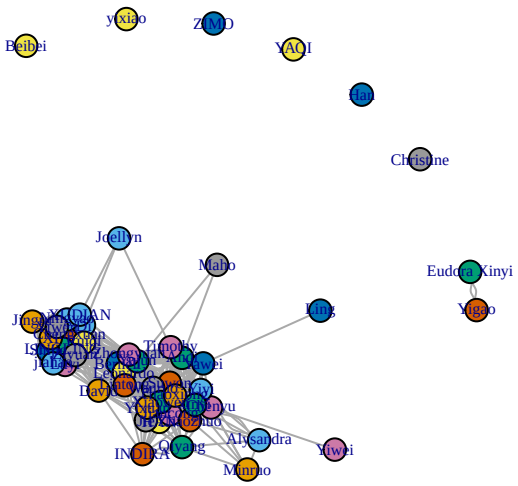
```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
D7 <- select(D6,FirstName,course)
#manage the course,that each student may type same course with diffterent way
D7 <- D7%>%filter(!course == "") %>% filter(course != "4050")
```

```
names(D7)<- c("student","course")
D7$course <- gsub(" ","",D7$course)
D7$course <- gsub("5026","HUDM5026",D7$course)
D7$course <- gsub("5126","HUDM5126",D7$course)
D7$course <- gsub("QMSS","G",D7$course)
D7$course <- gsub("GG","G",D7$course)
D7$course <- gsub("GGR","G",D7$course)
#drop hudk4050, since everyone takes hudk4050
D7 <- D7%>% filter(course != "HUDK4050")
#generate person-class matrix
D8<- mutate(D7,enrolled = 1)
D9 <- spread(D8,course,enrolled,0)
#generate person-person matrix
D10<-select(D9,-student)
D10 <- as.matrix(D10)
D11 <- t(D10)
D12<- D10 %*% D11
diag(D12)<-0
colnames(D12)<-D9$student
rownames(D12)<-D9$student
#plot igraph
g1 <-graph_from_adjacency_matrix(D12,mode="undirected")
V(g1)$label.cex <- seq(0.5,0.5,length.out=6)
```

```
## Warning in vattrs[[name]][index] <- value: number of items to replace is
## not a multiple of replacement length
```

```
plot(g1,layout=layout.fruchterman.reingold,,vertex.size = 10,edge.arrow.size=0.3,vertex.label = D9$stude
```



get betweeness centrality and max betweeness centrality person

```
peoplebetween=betweenness(g1)
maxpeoplebetween <-peoplebetween[peoplebetween == max(peoplebetween)]
maxpeoplebetween
```

```
##    Yujun
## 79.03208
```

get degree and people with max degree

```
degree <- degree(g1)
maxdegree <- degree[degree == max(degree)]
maxdegree
```

```
## Lintong
##      46
```

To Submit Your Assignment

Please submit your assignment by first "knitting" your RMarkdown document into an html file and then comit, push and pull request both the RMarkdown file and the html file.