Exercise 1

```
df <- read.csv("C:/Users/csg20/Downloads/ppl analytic/Connections.csv", header = TRUE)
colnames(df) <- gsub("\\.", "_", colnames(df))
colnames(df)
```

```
## [1] "First_Name"    "Last_Name"     "URL"           "Email_Address"
## [5] "Company"       "Position"      "Connected_On"
```

```
library(dplyr)
```

```
## Warning: package ’dplyr’ was built under R version 4.2.3
```

```
##
## Attaching package: ’dplyr’
```

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

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

```
library(tidyverse)
```

```
## Warning: package ’tidyverse’ was built under R version 4.2.2
```

```
## Warning: package ’ggplot2’ was built under R version 4.2.3
```

```
## Warning: package ’tibble’ was built under R version 4.2.3
```

```
## Warning: package ’tidyr’ was built under R version 4.2.2
```

```
## Warning: package ’readr’ was built under R version 4.2.2
```

```
## Warning: package ’purrr’ was built under R version 4.2.3
```

```
## Warning: package ’stringr’ was built under R version 4.2.2
```

```
## Warning: package ’forcats’ was built under R version 4.2.2
```

```
## Warning: package ’lubridate’ was built under R version 4.2.3
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.5.0     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(tidygraph)
```

```
## Warning: package 'tidygraph' was built under R version 4.2.3
```

```
##
## Attaching package: 'tidygraph'
##
## The following object is masked from 'package:stats':
##
##     filter
```

```r
library(ggraph)
```

```
## Warning: package 'ggraph' was built under R version 4.2.3
```

```r
library(igraph)
```

```
## Warning: package 'igraph' was built under R version 4.2.3
```

```
##
## Attaching package: 'igraph'
##
## The following object is masked from 'package:tidygraph':
##
##     groups
##
## The following objects are masked from 'package:lubridate':
##
##     %--%, union
##
## The following objects are masked from 'package:purrr':
##
##     compose, simplify
##
## The following object is masked from 'package:tidyr':
##
##     crossing
##
## The following object is masked from 'package:tibble':
##
```

```
##     as_data_frame
##
## The following objects are masked from ’package:dplyr’:
##
##     as_data_frame, groups, union
##
## The following objects are masked from ’package:stats’:
##
##     decompose, spectrum
##
## The following object is masked from ’package:base’:
##
##     union
```

```
library(readr)
```

```
df <- df %>%
  filter(!((is.na(First_Name) | First_Name == "") & (is.na(Last_Name) | Last_Name == "")))

df <- df %>%
  distinct(First_Name, Last_Name, .keep_all = TRUE)
```

```
count_by_employer <- df %>%
  group_by(Company) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count))

# Display the counts
print(count_by_employer)
```

```
## # A tibble: 307 x 2
##    Company                                               Count
##    <chr>                                                 <int>
##  1 "McGill University - Desautels Faculty of Management"    23
##  2 "McGill University"                                      16
##  3 ""                                                       11
##  4 "KPMG Canada"                                             7
##  5 "L’Oréal"                                                 7
##  6 "Desautels Capital Management"                            6
##  7 "CN"                                                      5
##  8 "Deloitte"                                                5
##  9 "TD"                                                      5
## 10 "BNP Paribas"                                             4
## # i 297 more rows
```

```
df %>%
  count()
```

```
##     n
## 1 449
```

```r
# Create nodes dataframe with unique individuals
nodes <- df %>%
  mutate(Label = paste(First_Name, substr(Last_Name, 1, 1), sep = " ")) %>%
  select(Label, Company) %>%
  distinct() %>%
  filter(!is.na(Label))
```

```r
# Create edges based on shared company affiliation
edges <- nodes %>%
  select(Label, Company) %>%
  distinct() %>%
  inner_join(nodes %>% select(Label, Company) %>% distinct(), by = "Company", suffix = c("_x", "_y")) %
  filter(Label_x != Label_y) %>%
  select(from = Label_x, to = Label_y) %>%
  distinct()
```

```
## Warning in inner_join(., nodes %>% select(Label, Company) %>% distinct(), : Detected an unexpected ma
## i Row 4 of ‘x‘ matches multiple rows in ‘y‘.
## i Row 13 of ‘y‘ matches multiple rows in ‘x‘.
## i If a many-to-many relationship is expected, set ‘relationship =
##   "many-to-many"‘ to silence this warning.
```

```r
# Ensure there are no NA values and that 'from' and 'to' columns are properly matched to node IDs
edges <- edges %>%
  filter(!is.na(from) & !is.na(to)) %>%
  left_join(nodes, by = c("from" = "Label")) %>%
  left_join(nodes, by = c("to" = "Label"))
```

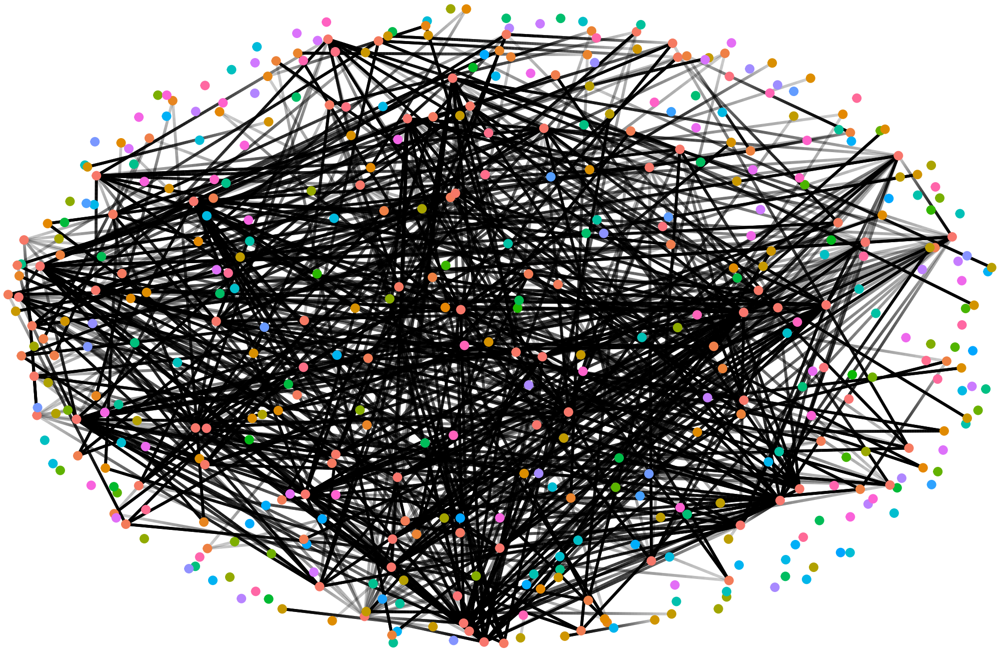
```r
tidy_graph <- tbl_graph(nodes = nodes, edges = edges, directed = FALSE)

tidy_graph <- tidy_graph %>%
  activate(nodes) %>%
  mutate(community = as.factor(group_louvain()))

# Create the graph visualization
g <- ggraph(tidy_graph, layout = 'kk') +
  geom_edge_link(aes(alpha = stat(index)), show.legend = FALSE) +
  geom_node_point(aes(colour = community), show.legend = FALSE, size = 1) +
  theme_graph()

g
```

```
## Warning: ‘stat(index)‘ was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(index)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



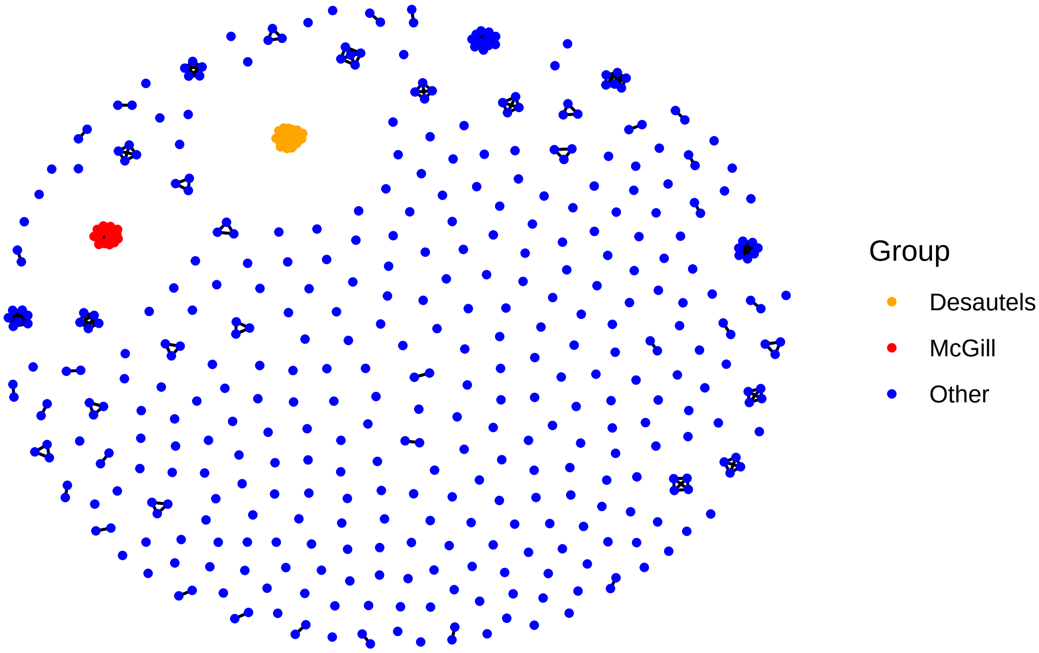
```r
nodes <- df %>%
  mutate(Label = paste(First_Name, substr(Last_Name, 1, 1), sep = " ")) %>%
  select(Label, Company) %>%
  distinct() %>%
  filter(!is.na(Label)) %>%
  mutate(group = case_when(
    Company == "McGill University - Desautels Faculty of Management" ~ "Desautels",
    Company == "McGill University" ~ "McGill",
    TRUE ~ "Other"))


tidy_graph <- tbl_graph(nodes = nodes, edges = edges, directed = FALSE)

tidy_graph <- tidy_graph %>%
  activate(nodes) %>%
  mutate(community = as.factor(group_louvain()))

# Create the graph visualization
g <- ggraph(tidy_graph, layout = 'fr') +
  geom_edge_link() +
  geom_node_point(aes(color = group), size = 1) +
  theme_graph()+
  scale_color_manual(values = c("Desautels" = "orange", "McGill" = "red", "Other" = "blue")) +
  theme_graph(base_family="sans") +
  labs(color = "Group")
```

```
# Print the graph
g
```



The clustering of Desautels and McGill nodes indicates a dense network of connections within university and faculty, which is expected given I am studying at Desautels at McGill. The ‘Other’ category represents connections from different facets such as past educational institutions, colleagues, friends, or industry professionals not directly associated with McGill or Desautels. The Desautels cluster is small and isolated, indicating little to no direct interaction with nodes from other groups. In contrast, the McGill cluster is closer to the ‘Other’ nodes, suggesting some level of interaction or connection between the McGill group and the ‘Other’ group.