Ex 1

Heidi Al Wakeel

2023-03-14

Check the data

```
df <-  read.csv("C:/Users/Heidi Al Wakeel/Downloads/Basic_LinkedInDataExport_03-09-2023/Connections.csv")
df <- df %>%
  select(-c("Email.Address"))
```

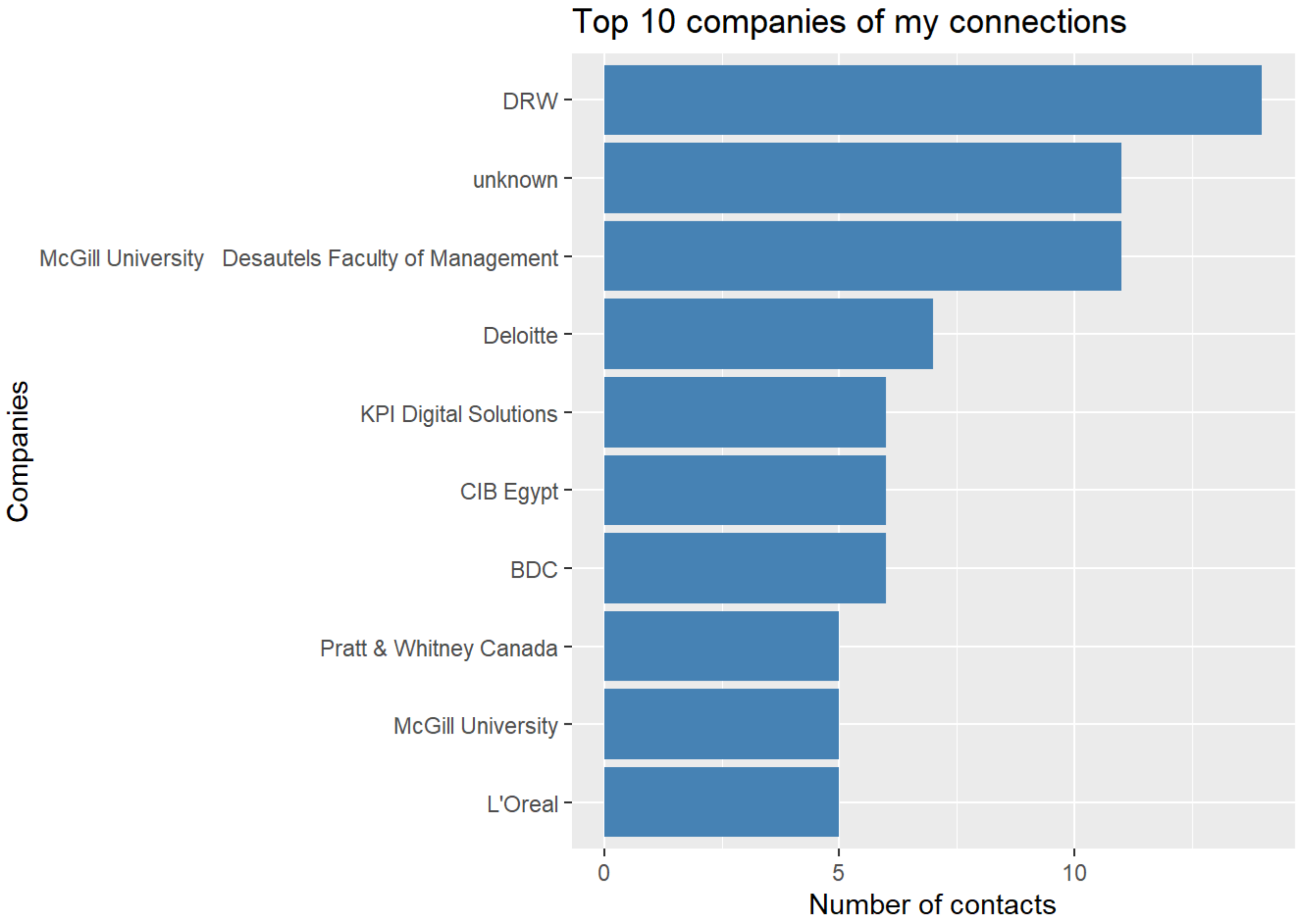
Standardize the name of the companies

There are still some duplicates with inc. or canada. A more systematic way by using the companies name resemblance could be used but that’s not the goal of this project and most of the duplicates were fixed by doing these simple fix.

```
# Lower case company name
df <-
  df %>%
  mutate(company = tolower(Company))  # lower case
# If no company, write "unknown"
df <-
  df %>%
  mutate(Company = replace_na(Company, "None")) %>%
  mutate(Company = replace(Company, Company=="", "unknown"))
# Remove accents in the column
df$Company <- stri_trans_general(str=df$Company, id="Latin-ASCII")
# Replace everything starting with McGill by just McGill
df <- df %>%
  mutate(company = replace(Company, str_detect(Company, "mcgill"), "mcgill"))
# Remove "-" and replace with space
df <- df %>%
  mutate(Company = str_replace(Company, "-", " "))
# PRIVACY
#df %>% head(10)
```

Get the count of contacts by company

```
count <- df %>%
  group_by(Company) %>%
  count() %>%
  arrange(desc(n))

count %>% arrange(desc(n)) %>% head(10) %>%
  ggplot(aes(y = reorder(Company,n), x=n))+
  geom_col(fill="steelblue") +
  labs(
    x = "Number of contacts",
    y = "Companies",
    title = "Top 10 companies of my connections"
  )
```



```
count
```

```
## # A tibble: 295 × 2
## # Groups:   Company [295]
##    Company                                            n
##    <chr>                                          <int>
##  1 DRW                                               14
##  2 McGill University   Desautels Faculty of Management    11
##  3 unknown                                           11
##  4 Deloitte                                           7
##  5 BDC                                                6
##  6 CIB Egypt                                          6
##  7 KPI Digital Solutions                              6
##  8 L'Oreal                                            5
##  9 McGill University                                  5
## 10 Pratt & Whitney Canada                             5
## # … with 285 more rows
```

Get the total count

```
total_count = sum(count$n)
print(c("Total connections = ", total_count))
```

```
## [1] "Total connections = " "400"
```

Create the graph

Create a column with the first and last name

```
df <- df %>%
  unite(name, c("First.Name", "Last.Name"))
```

Remove the unknown company contacts from the network

```
df <- df %>% filter(Company!="unknown")
```

Create the nodes

```
nodes <- df %>% select(c("name", "Company"))
nodes <- nodes %>% rowid_to_column("id")
```

Create the edges

Left join the id of the contact’s name with the same company name

```
edges <- df %>% select(c(name, Company)) %>%
  left_join(nodes %>% select(c(id,name)), by = c("name"="name"))
```

```
## Warning in left_join(., nodes %>% select(c(id, name)), by = c(name = "name")): Each row in `x
` is expected to match at most 1 row in `y`.
## ℹ Row 266 of `x` matches multiple rows.
## ℹ If multiple matches are expected, set `multiple = "all"` to silence this
##   warning.
```

```
edges <- edges %>% left_join(edges, by = "Company", keep=FALSE) %>%
  select(c("id.x", "id.y", "Company")) %>%
  filter(id.x!=id.y) # remove the connections between itself
```

```
## Warning in left_join(., edges, by = "Company", keep = FALSE): Each row in `x` is expected to
match at most 1 row in `y`.
## ℹ Row 2 of `x` matches multiple rows.
## ℹ If multiple matches are expected, set `multiple = "all"` to silence this
##   warning.
```

```
colnames(edges) <- c("x", "y", "Company")
edges %>% head(10)
```

```
##    x   y                                             Company
## 1  2  41                                   McGill University
## 2  2 116                                   McGill University
## 3  2 162                                   McGill University
## 4  2 386                                   McGill University
## 5  3  39 McGill University   Desautels Faculty of Management
## 6  3  43 McGill University   Desautels Faculty of Management
## 7  3  44 McGill University   Desautels Faculty of Management
## 8  3  45 McGill University   Desautels Faculty of Management
## 9  3  61 McGill University   Desautels Faculty of Management
## 10 3  79 McGill University   Desautels Faculty of Management
```

Create the graph

```
library(tidygraph)
```

```
##
## Attaching package: 'tidygraph'
## The following object is masked from 'package:igraph':
##
##     groups
```

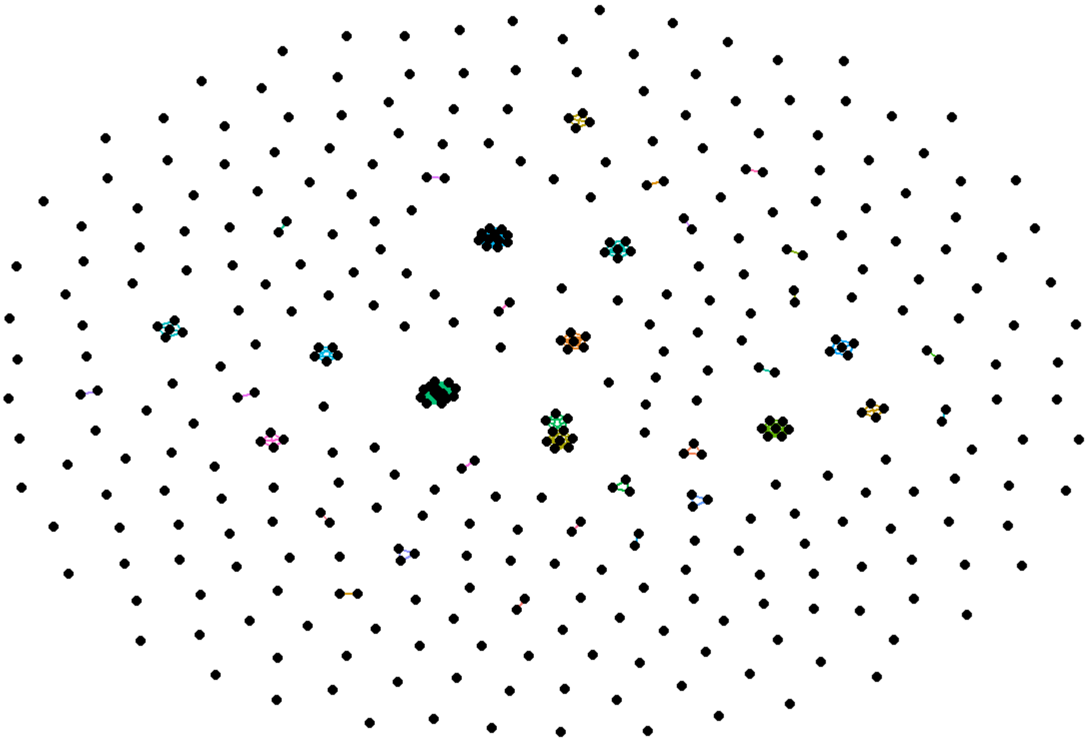
```
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(ggraph)
graph <- tbl_graph(edges = edges, nodes=nodes, directed = FALSE)
```

Plot the resulting full graph

```
ggraph(graph, layout = "graphopt") +
  geom_edge_link(aes(color = Company), show.legend = FALSE) +
  geom_node_point()+
  theme_graph()
```

```
## Warning: Using the `size` aesthetic in this geom was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` in the `default_aes` field and elsewhere instead.
```



Now, I wanted to take a deeper dive:

```
connections <- read.csv("C:/Users/Heidi Al Wakeel/Downloads/Basic_LinkedInDataExport_03-09-2023/
Connections.csv")



connections <- na.omit(connections)



attach(connections)
```

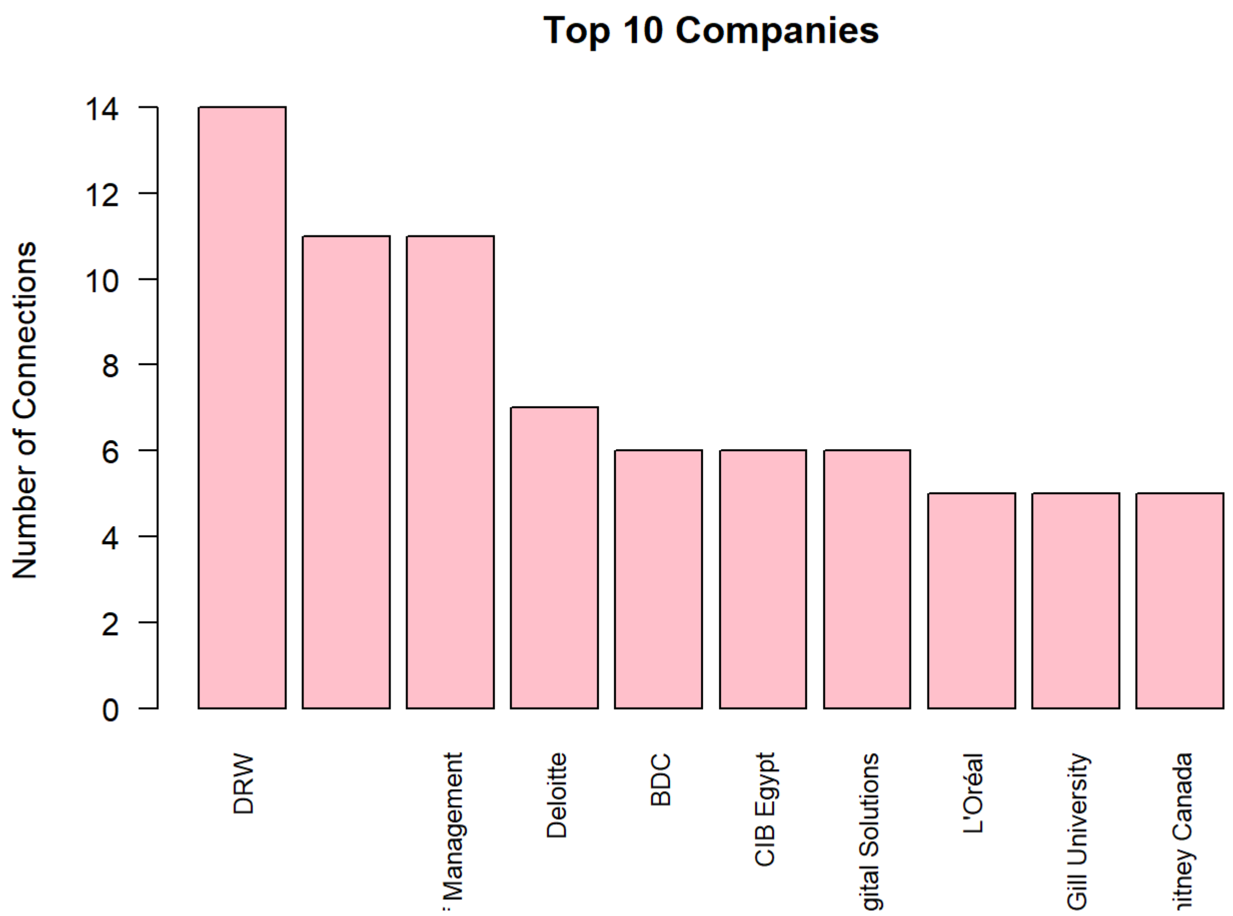
```
## The following object is masked from package:ggplot2:
##
##     Position
```

```
# Create a table with
connections$name =  paste(connections$First.Name, substr(connections$Last.Name, start = 1, stop
= 1), sep = " ")

connections = connections[, c("name", "Company","Position", "Connected.On")]

# create a frequency table
freq_table = table(connections$Company)
freq_table = sort(freq_table, decreasing = TRUE)
first10= head(freq_table, n = 10)


# Display the list as a table
knitr::kable(first10, col.names = c("Company", "Connections"))
```

Company	Connections
DRW	14
	11
McGill University - Desautels Faculty of Management	11
Deloitte	7
BDC	6
CIB Egypt	6
KPI Digital Solutions	6
L’Oréal	5
McGill University	5
Pratt & Whitney Canada	5

Top 10 Companies and their frequencies

```
# create a bar chart of the frequency table
barplot(first10, main = "Top 10 Companies",
         ylab = "Number of Connections",
        col = "Pink", las = 2, cex.names = 0.8)
```



Creating nodes

```
library(tidyverse)

people <- connections %>%
  distinct(name) %>%
  rename(label = name)

companies <- connections %>%
  distinct(Company) %>%
  rename(label = Company)

nodes <- full_join(people, companies, by = "label")
nodes <- rowid_to_column(nodes, "id")
head(nodes)
```

```
##   id      label
## 1  1     Tema H
## 2  2   Sameer G
## 3  3    Fatih N
## 4  4  Edouard S
## 5  5     Taju P
## 6  6 Adrianna E
```

Creating edges

```
#### Creating edges

edges <- connections[, c("name", "Company")]

edges <- edges %>%
  left_join(nodes, by = c("name" = "label")) %>%
  rename(from = id)

edges <- edges %>%
  left_join(nodes, by = c("Company" = "label")) %>%
  rename(to = id)

edges <- unique(select(edges, from, to))
head(edges)
```

```
##   from  to
## 1    1 378
## 2    2 379
## 3    3 380
## 4    4 381
## 5    5 382
## 6    6 383
```

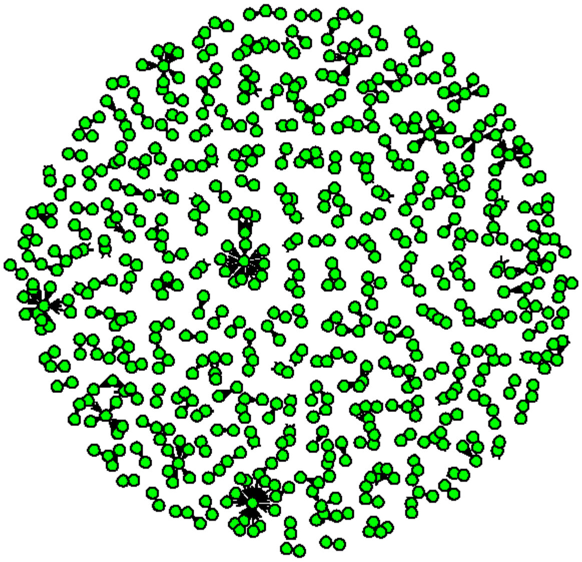
Graph using the network library

```
## Building network
library(network)
```

```
##
## 'network' 1.18.1 (2023-01-24), part of the Statnet Project
## * 'news(package="network")' for changes since last version
## * 'citation("network")' for citation information
## * 'https://statnet.org' for help, support, and other information
```

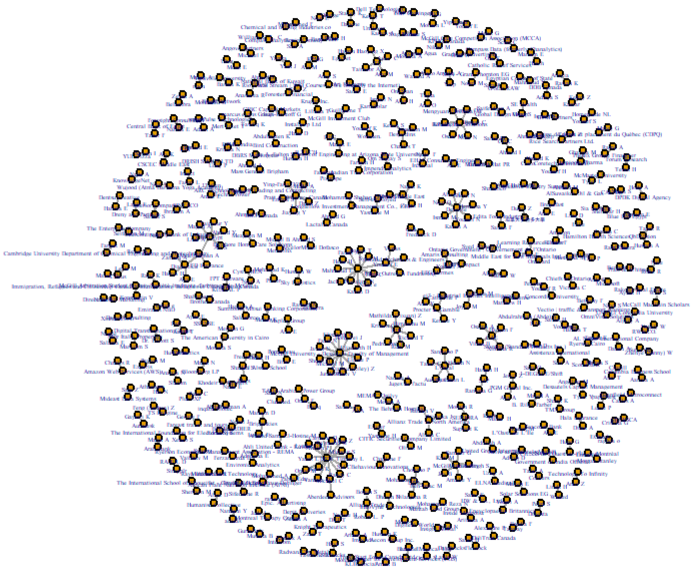
```
##
## Attaching package: 'network'
## The following objects are masked from 'package:igraph':
##
##     %c%, %s%, add.edges, add.vertices, delete.edges, delete.vertices,
##     get.edge.attribute, get.edges, get.vertex.attribute, is.bipartite,
##     is.directed, list.edge.attributes, list.vertex.attributes,
##     set.edge.attribute, set.vertex.attribute
```

```
routes_network <- network(edges,
                          vertex.attr = nodes,
                          matrix.type = "edgelist",
                          ignore.eval = FALSE)
plot(routes_network, vertex.cex = 1,vertex.col="green")
```



Graph using igraph

```
## igraph

library(igraph)
routes <- graph_from_data_frame(d = edges,
                                       vertices = nodes,
                                       directed = TRUE)


plot(routes,
     vertex.size = 3,
     vertex.label.cex = 0.2,
     edge.arrow.size = 0.01, vertex.col="green")
```



```
library(dplyr)

connections_filtered <- connections %>%
  group_by(Company) %>%
  filter(n() > 1) %>%
  ungroup()
```

Recreating nodes for companies with more than 1 connection

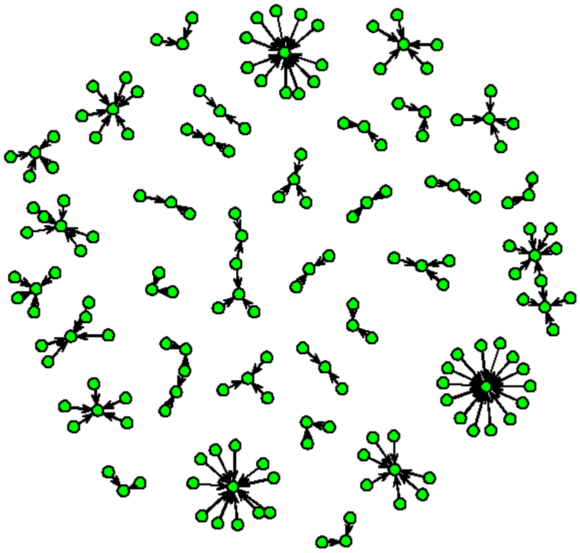
```
people <- connections_filtered %>%
  distinct(name) %>%
  rename(label = name)

companies <- connections_filtered %>%
  distinct(Company) %>%
  rename(label = Company)

nodes <- full_join(people, companies, by = "label")
nodes <- rowid_to_column(nodes, "id")
```

Recreating edges

```
#### Creating edges

edges <- connections_filtered[, c("name", "Company")]

edges <- edges %>%
  left_join(nodes, by = c("name" = "label")) %>%
  rename(from = id)

edges <- edges %>%
  left_join(nodes, by = c("Company" = "label")) %>%
  rename(to = id)

edges <- unique(select(edges, from, to))
```

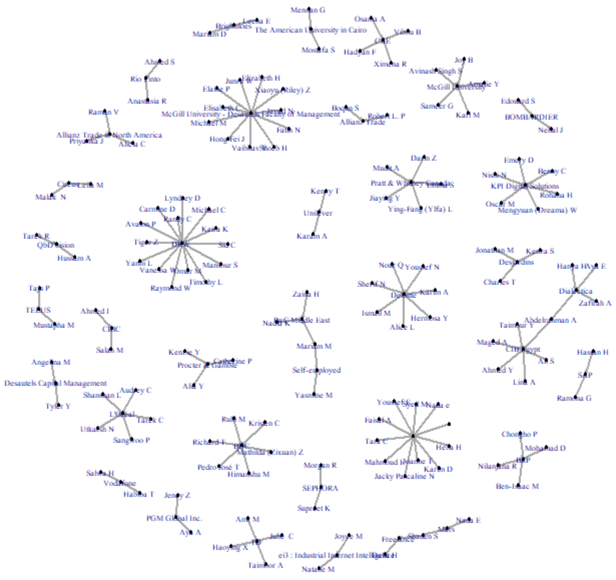
Graph using the network library

```
## Building network
library(network)

routes_network <- network(edges,
                          vertex.attr = nodes,
                          matrix.type = "edgelist",
                          ignore.eval = FALSE)
plot(routes_network, vertex.cex = 1, vertex.col="green")
```



Graph using igraph

```
## igraph

library(igraph)
routes <- graph_from_data_frame(d = edges,
                                       vertices = nodes,
                                       directed = TRUE)


plot(routes,
     vertex.size = 1,
     vertex.label.cex = 0.25,
     edge.arrow.size = 0.05
     )
```


Menhan G
Leena E
Brightskies
Mariam D
The American University in Cairo
Mostafa S
Osama A
Vibhu B
GE
Hadyan F
Ximena R
Joy B
Avinash Singh S
Amelie Y
McGill University
Sameer G
Kari M
Edouard S
BOMBARDIER
Nehal J
Ahmed S
Rio Tinto
Anastasia R
Junaid W
Elizabeth H
Elaine P
Xiaoyu (Riley) Z
Elisabeth C
Junaid N
McGill University - Desautels Faculty of Management
Michael M
Faith N
Hong Fei J
Vaibhav S
Shoeb H
Bogdan S
Robert L. P
Allianz Trade
Raman V
Allianz Trade in North America
Priyanka J
Alicia C
Darin Z
Maria A
Pratt & Whitney Canada
Emma S
Jiaying Y
Ying-Fang (Yifa) L
Emery D
Nishi N
Benny C
KPI Digital Solutions
Rohana H
Oscar M
Mengyuan (Dreama) W
Cirrus
Leila M
Malak N
Kenny T
Unilever
Karim A
Lyndsey D
Carmine D
Michael C
Randy C
Avalon P
Keith K
Tigar Z
DHW
Sid C
Yanni L
Mansour S
Vanessa W
Omar M
Timothy L
Raymond W
Tarek R
QbD Vision
Hussam A
Jonathan M
Kendra S
Desjardins
Charles T
Hanya HAya E
Dialectica
Zafirah A
Noor Q
Yousef N
Sherif N
Deloitte
Karim A
Ismail M
Hermosa Y
Alice L
Taja P
TELUS
Mustapha M
Ahmed I
CIBC
Salah M
Zaina H
Nadia K
P&G Middle East
Mariam M
Self-employed
Yasmine M
Abdelrahman A
Taimour Y
Maged A
CIB Egypt
Ali S
Ahmed Y
Lina A
Hassan H
SAP
Ramona G
Kendie Y
Catherine P
Procter & Gamble
Alia Y
Angelina M
Desautels Capital Management
Tyler Y
Audrey C
Shannan L
L'Oreal
Tarek C
Utkarsh N
Sangwoo P
Youssef K
Syed M
Nada e
Faisal A
Tara C
Heba H
Mahmoud I
Joanne T
Karim D
Jacky Pascaline N
Ruth M
Kristen C
Richard T
BMO
Mathilda (Zixuan) Z
Pedro José T
Himanshu M
Chongho P
Mohamad D
Nilanjana R
BRP
Ben-Isaac M
Morgan R
SEPHORA
Supreet K
Salma H
Vodafone
Habiba T
Jenny Z
PGM Global Inc.
Aya A
Anil M
Julie C
IBM
Haoying X
Taimoor A
Joyce M
ei3 : Industrial Internet Intelligence
Natalie M
Nada E
MBS
Shaden S
Freelance
Dana H

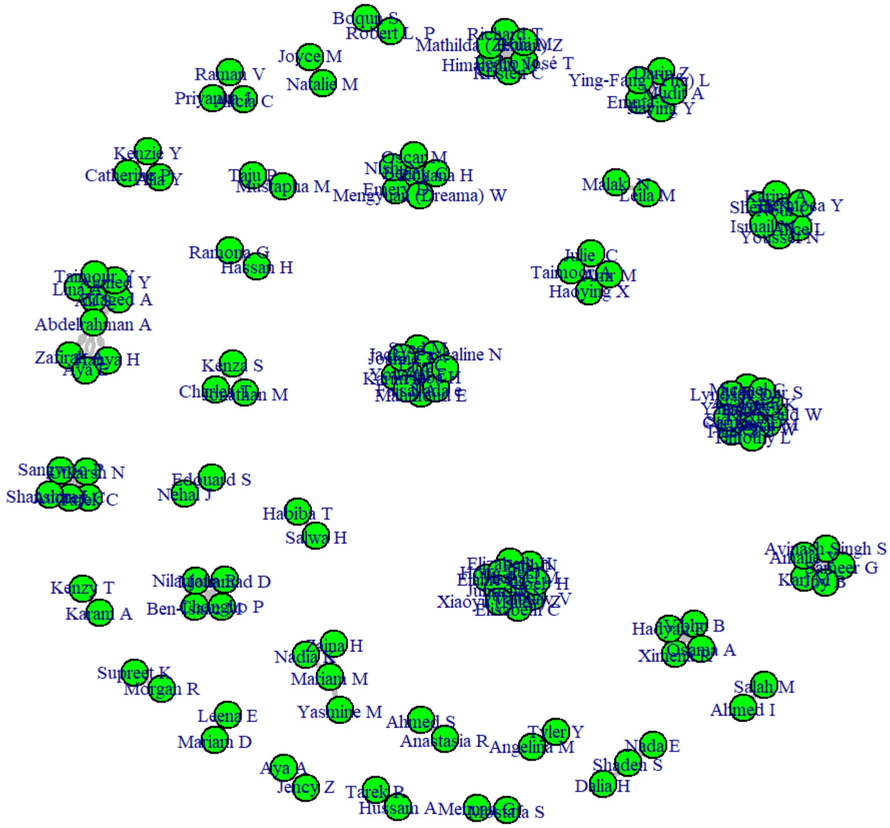
Final graph with no companies

```
# Filter connections to only include companies with 2 or more employees

contact_count <- connections %>%
  group_by(Company) %>%
  summarize(count = n())

Connections <- connections %>%
  inner_join(contact_count, by = "Company") %>%
  filter(count >= 2) %>%
  select(name, Company)

# Create nodes dataframe using tidygraph
nodes <- Connections %>%
  mutate(label = name) %>%
  distinct(label) %>%
  as_tibble() %>%
  select(label)

# Create edges dataframe using tidygraph
edges <- Connections %>%
  left_join(connections, by = "Company") %>%
  filter(name.x != name.y) %>%
  mutate(from = name.x,
         to = name.y) %>%
  select(from, to)
```

```
## Warning in left_join(., connections, by = "Company"): Each row in `x` is expected to match at
most 1 row in `y`.
## ℹ Row 1 of `x` matches multiple rows.
## ℹ If multiple matches are expected, set `multiple = "all"` to silence this
##   warning.
```

```
# Create graph using igraph
graph <- graph_from_data_frame(edges, vertices = nodes, directed = FALSE)
```

```
par(mar = rep(1, 4))
options(repr.plot.width = 10, repr.plot.height = 10)
plot(graph, vertex.size = 7, vertex.color = "green", vertex.label.cex = 0.6, edge.color = "gra
y", edge.width = 2, edge.length=30, vertex.dist = 50)
```


Boqun S
Robert L. P
Richard T
Mathilda
MZ
Himani
José T
Krister C
Joyce M
Natalie M
Raman V
Priyanka
Alicia C
Ying-Fang
Darin Z
Judit A
Emma
Jiaying Y
Kenzie Y
Catherine P
Y
Taju P
Mustapha M
Oscar M
Ruqayyana H
Emery D
Mengyuan (Dreama) W
Malak N
Leila M
Karim A
Shehab
Y
Ismail
L
Youssef N
Ramona G
Hassan H
Julie C
Taimoor A
M
Haoying X
Taimour Y
Lina A
Khaled Y
Amjed A
Abdelrahman A
Zafirah
Hanya H
Aya E
Kenza S
Charles T
Jonathan M
Syed M
Adeline N
Jacque
Johanne
Karim
Yousef
E H
Faisal
Nada
Mahmoud E
Lynda
Manar
S
K
W
W
W
Timothy L
Sangyoon P
Utkarsh N
Shamim
Ashraf
Tarek C
Edouard S
Nehal J
Habiba T
Salwa H
Avinash Singh S
Amalie Y
Sameer G
Karlow B
Elizabeth N
M
H
Julian
Xiaoyi
Z
V
Elizabeth C
Kenzy T
Karam A
Nilanjana
Ahmad D
Ben-Isaac
Zhonghe P
Hadya
B
Osama A
Ximena R
Supreet K
Morgan R
Zaina H
Nadia K
Mariam M
Yasmine M
Salah M
Ahmed I
Leena E
Mariam D
Ahmed S
Anastasia R
Tyler Y
Angelina M
Nada E
Shaden S
Dalia H
Aya A
Jency Z
Tarek R
Hussam A
Mostafa S