MSD final project

```r
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------------
```

```
## v ggplot2 2.2.1       v purrr   0.3.0
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.1       v stringr 1.3.1
## v readr   1.1.1       v forcats 0.3.0
```

```
## Warning: package 'tibble' was built under R version 3.4.4
```

```
## Warning: package 'tidyr' was built under R version 3.4.4
```

```
## Warning: package 'purrr' was built under R version 3.4.4
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
## -- Conflicts ---------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(modelr)
```

```
## Warning: package 'modelr' was built under R version 3.4.4
```

```r
library(ggplot2)
library(igraph)
```

```
## Warning: package 'igraph' was built under R version 3.4.4
```

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

```
## The following object is masked from 'package:modelr':
##
##     permute
```

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

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

```
##     union
```

```r
hist_edgelist = read.table( "Dataset 5. History_edgelist.txt", header = TRUE)
hist_vertex = read.table(file = 'Dataset 6. History_vertexlist.txt', sep = '\t', header = TRUE)

head(hist_edgelist)
```

```
##   u v  rank gender
## 1 1 1 Assoc      F
## 2 1 1  Full      F
## 3 1 1  Full      F
## 4 1 1  Full      M
## 5 1 1  Full      M
## 6 1 1  Full      M
```

```r
head(hist_vertex)
```

```
##   u   pi USN2009 NRC2010    Region           institution
## 1 1 1.54       5       1 Northeast    Harvard University
## 2 2 2.41       1      12 Northeast       Yale University
## 3 3 4.80       1      14 West                UC Berkeley
## 4 4 5.16       1       1 Northeast  Princeton University
## 5 5 5.45       1       9 West        Stanford University
## 6 6 6.19       5       4 Midwest   University of Chicago
```

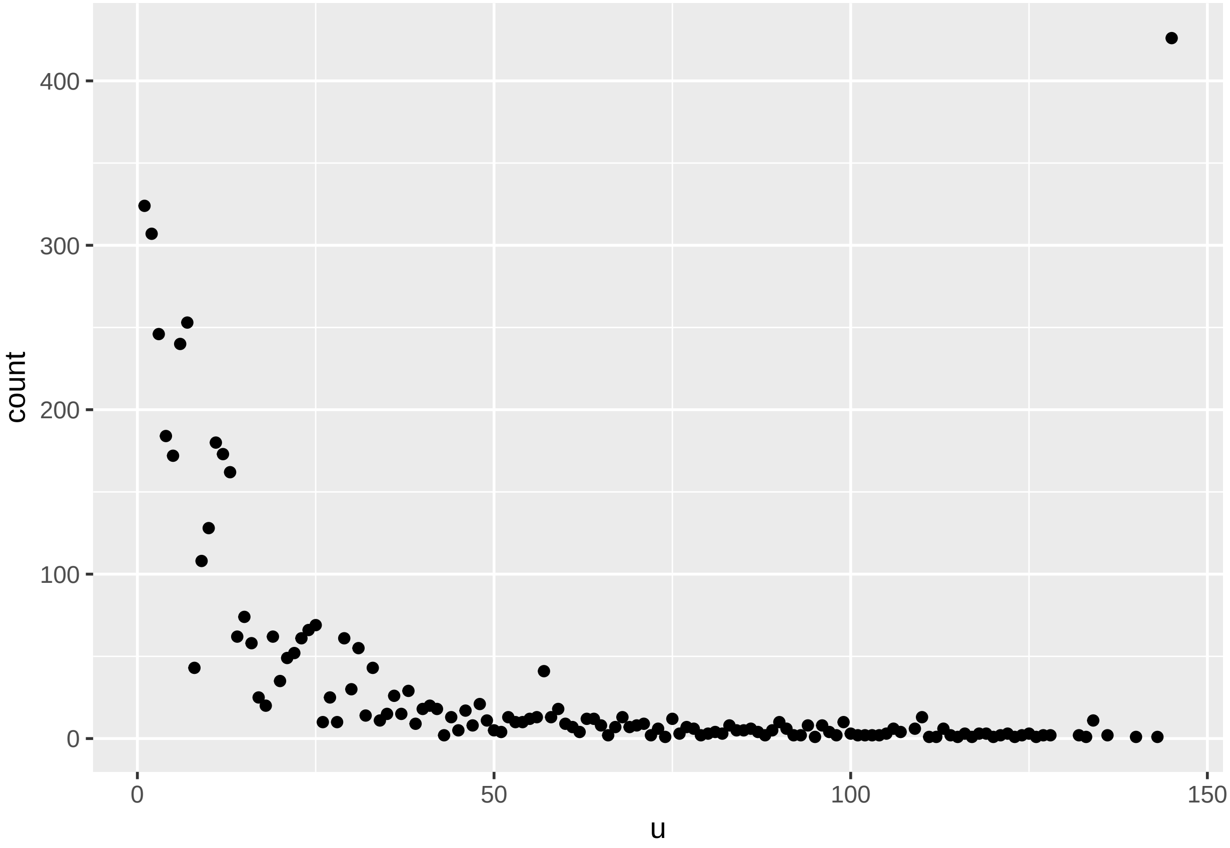
```r
employee_counts = hist_edgelist %>%
  group_by( u ) %>%
  summarize( count = n()) %>%
  ungroup() %>%
  left_join( hist_vertex, by = "u") %>%
  select(u, count, institution)

grad_counts = hist_edgelist %>%
  group_by( v ) %>%
  summarize( count = n() ) %>%
  ungroup()


head(  employee_counts )
```

```
## # A tibble: 6 x 3
##       u count institution
##   <int> <int> <fct>
## 1     1   324 Harvard University
## 2     2   307 Yale University
## 3     3   246 UC Berkeley
## 4     4   184 Princeton University
## 5     5   172 Stanford University
## 6     6   240 University of Chicago
```

```r
head(  grad_counts )
```

```
## # A tibble: 6 x 2
##       v count
##   <int> <int>
## 1     1    45
## 2     2    62
## 3     3    47
## 4     4    60
## 5     5    49
## 6     6    46
```

```
employee_counts %>%
  ggplot(aes(x = u, y = count)) +
  geom_point()
```



```
tail( employee_counts )
```

```
## # A tibble: 6 x 3
##       u count institution
##   <int> <int> <fct>
## 1   133     1 Southern Illinois University, Carbondale
## 2   134    11 Southern Baptist Theological Seminary
## 3   136     2 Saint Louis University
## 4   140     1 University of Memphis
## 5   143     1 Oklahoma State University
## 6   145   426 All others
```

```
#getting rid of row 145, "All Others" Doesn't provide much info
employee_counts = employee_counts %>%
  filter(u != 145)
```

Making a network of weighted edges

```
hist_weighted_edgelist = hist_edgelist %>%
  group_by(v, u) %>%
  summarize( count = n()) %>%
  ungroup() %>%
  left_join( hist_vertex, by = c('v'= 'u')) %>%
  select(v, u, count, institution)


tail( hist_weighted_edgelist )
```

```
## # A tibble: 6 x 4
##       v     u count institution
##   <int> <int> <int> <fct>
## 1   144    80     1 Middle Tennessee State University
## 2   144    88     1 Middle Tennessee State University
## 3   144    90     1 Middle Tennessee State University
## 4   144   100     1 Middle Tennessee State University
## 5   144   127     2 Middle Tennessee State University
## 6   144   136     1 Middle Tennessee State University
```

Filtering to just the schools with a count > 100 so I can make a graph to just look at the network

```
ids = employee_counts %>%
  filter(count > 100) %>%
  select(u)





smaller = hist_weighted_edgelist %>%
  filter( u %in% ids$u , v %in% ids$u)


nrow( smaller )
```

```
## [1] 126
```

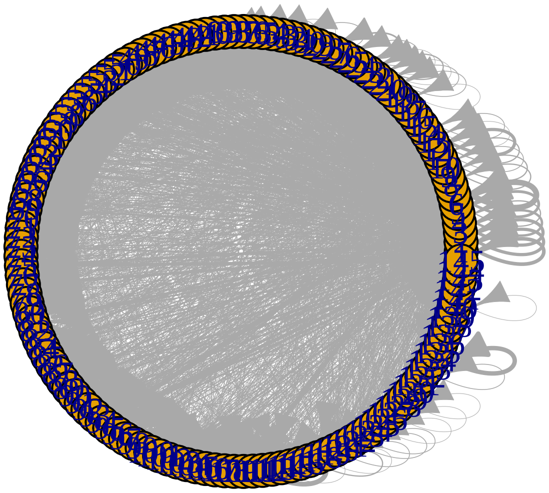
```
nrow(hist_edgelist)
```

```
## [1] 4538
```

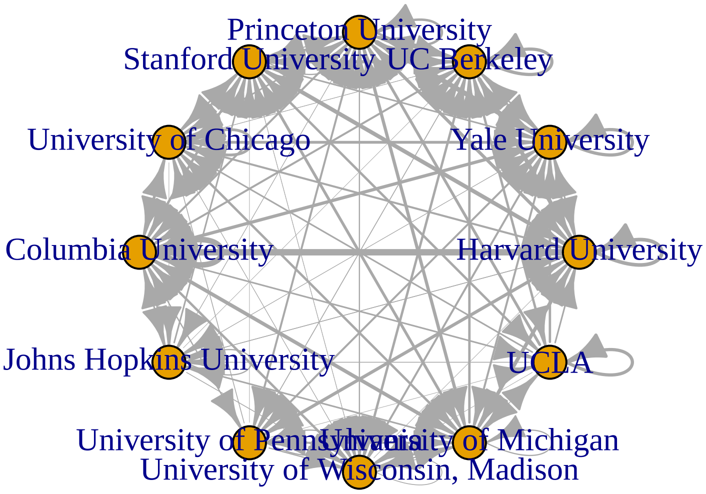
```
ids
```

```
## # A tibble: 12 x 1
##        u
##    <int>
##  1     1
##  2     2
##  3     3
##  4     4
##  5     5
##  6     6
##  7     7
##  8     9
##  9    10
## 10    11
## 11    12
## 12    13
```

```
graph = hist_weighted_edgelist %>%
  graph_from_data_frame(directed = TRUE)

smaller_graph = smaller %>%
   graph_from_data_frame(directed = TRUE)

plot(
  graph, vertex_size = 1, edge.width=E(graph)$count/5, layout =  layout_in_circle(graph, order = V(grap
    )
  )
```



```
plot(smaller_graph, vertex_size = 2 ,edge.width=E(smaller_graph)$count/5,
     layout = layout_in_circle(smaller_graph, order = V(smaller_graph)),
     vertex.label = unique( E(smaller_graph)$institution )
     )
```



```
#For ideas of looking into how the networks change when filtering for these values

hist_edgelist %>%
  group_by(rank) %>%
  summarize( count = n() )
```

```
## # A tibble: 3 x 2
##   rank  count
##   <fct> <int>
## 1 Assoc  1609
## 2 Asst    844
## 3 Full   2085
```