Exercise 4

Utkarsh Pal

2023-04-04

Centrality and Efficiency

Loading the Necessary Packages

```
library(readr)
```

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

```
library(arrow)
```

```
## Warning: package 'arrow' was built under R version 4.2.2
```

```
##
## Attaching package: 'arrow'
```

```
## The following object is masked from 'package:utils':
##
##     timestamp
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
##
## 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(tidyr)
library(igraph)
```

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

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

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

```
## 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(tidygraph)
```

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

```
##
## Attaching package: 'tidygraph'
```

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

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

```
library(tibble)
```

```
##
## Attaching package: 'tibble'
```

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

Importing the Data from Exercise 3

```
applications = read_parquet("C:\\Users\\Utkarsh\\Desktop\\app_data_clean.parquet")
edges = read_csv("C:\\Users\\Utkarsh\\Desktop\\edges_sample.csv")
```

```
## Rows: 32906 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): application_number
## dbl  (2): ego_examiner_id, alter_examiner_id
## date (1): advice_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
edges = na.omit(edges)
```

Creating the Network

```
edges = edges %>%
  select(ego_examiner_id, alter_examiner_id, application_number)

nodes = edges %>%
  pivot_longer(cols = c('ego_examiner_id', 'alter_examiner_id')) %>%
  filter(name %in% c('ego_examiner_id', 'alter_examiner_id')) %>%
  distinct(value) %>%
  select(name = value)

network = graph_from_data_frame(edges, directed = TRUE, vertices=nodes) %>%
  as_tbl_graph()
```

Getting the Centrality Figures and Joining it with the Applications DF

```
network = network %>%
  mutate(degree = centrality_degree(),
         betweenness = centrality_betweenness(),
         closeness = centrality_closeness())

centralities = network %>%
  as_tibble() %>%
  rename(examiner_id = name)
```

```
centralities$closeness = ifelse(is.na(centralities$closeness), 0, centralities$closeness)
centralities$examiner_id = as.numeric(centralities$examiner_id)

applications = applications %>%
  left_join(centralities, on = 'examiner_id')
```

```
## Joining, by = "examiner_id"
```

```
# Removing Observations from Applications DF that do not have Centrality Figures
applications2 = applications[complete.cases(applications[, c("degree", "betweenness", "closeness")]), ]
```

Adding the 'app_proc_time' variable

```
applications2 = applications2 %>%
  mutate(app_proc_time = ifelse(is.na(patent_issue_date),
                                abandon_date - filing_date,
                                patent_issue_date - filing_date)) %>%
  select(app_proc_time, examiner_workgroup, gender, race, tenure_days, degree, betweenness, closeness) %
  mutate(examiner_workgroup = substr(examiner_workgroup, start = 1, stop = 2)) %>%
  rename(examiner_class = examiner_workgroup)
```

Further Data Transformation

```
# Converting examiner_class into a categorical variable
applications2$examiner_class = factor(applications2$examiner_class)
applications2 = subset(applications2, app_proc_time > 0)
# Standardizing numeric variables
applications2$tenure_days_std = scale(applications2$tenure_days)
applications2$degree_std = scale(applications2$degree)
applications2$betweenness_std = scale(applications2$betweenness)
applications2$closeness_std = scale(applications2$closeness)
applications2$app_proc_time_std = scale(applications2$app_proc_time)
applications2 <- na.omit(applications2)
```

Regression

```
m1 <- lm(app_proc_time_std ~ examiner_class + gender + race + tenure_days_std +
           degree_std + betweenness_std + closeness_std, data = applications2)
summary(m1)
```

```
##
## Call:
## lm(formula = app_proc_time_std ~ examiner_class + gender + race +
##     tenure_days_std + degree_std + betweenness_std + closeness_std,
##     data = applications2)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5472 -0.6651 -0.1673  0.4618  7.9241
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.083121   0.003102 -26.797  < 2e-16 ***
## examiner_class17 -0.048936   0.002795 -17.506  < 2e-16 ***
## examiner_class21  0.300803   0.003264  92.156  < 2e-16 ***
## examiner_class24  0.408552   0.003667 111.426  < 2e-16 ***
## gendermale       -0.046347   0.002316 -20.008  < 2e-16 ***
## raceblack        -0.070853   0.005806 -12.203  < 2e-16 ***
## raceHispanic      0.057725   0.007384   7.818 5.37e-15 ***
## raceother         0.297666   0.030459   9.773  < 2e-16 ***
## racewhite         0.010715   0.002438   4.395 1.11e-05 ***
## tenure_days_std  -0.027314   0.001070 -25.516  < 2e-16 ***
## degree_std        0.011986   0.001057  11.339  < 2e-16 ***
## betweenness_std   0.027989   0.001046  26.770  < 2e-16 ***
## closeness_std    -0.002874   0.001054  -2.728  0.00637 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9811 on 906601 degrees of freedom
## Multiple R-squared:  0.03754,    Adjusted R-squared:  0.03753
## F-statistic:  2947 on 12 and 906601 DF,  p-value: < 2.2e-16
```

The regression output above tells us that:

- There is a highly statistically significant negative correlation between examiners being in class 17 and the application processing time compared to the expected time it takes examiners in class 16.
- There is a highly statistically significant positive correlation between examiners being in class 21 & 24 and the application processing time compared to the expected time it takes examiners in class 16.
- There is a highly statistically significant negative correlation between examiners being males and the application processing time compared to the expected time it takes female examiners.
- There is a highly statistically significant negative correlation for black examiners and positive correlation for examiners of Hispanic, White, and Other race in the context of race and the application processing time compared to the expected time it takes Asian examiners.
- There is a highly statistically significant negative correlation between an examiner's tenure days and the application processing time. In other words, the longer an examiner has been in the workforce, the less time they are taking to process applications.
- There is a highly statistically significant positive correlation between degree & betweenness centralities and the application processing time. A possible explanation might be that examiners who are more connected to other examiners (i.e., high degree centrality) or who are more central in the network (i.e., high betweenness centrality) may have more responsibilities or may be more involved in complex cases, which could contribute to longer processing times. On the other hand, closeness centrality is negatively correlated with application processing time, indicating that examiners who have higher values of closeness centrality tend to process applications more quickly. A possible explanation may be that examiners with high closeness centrality have a shorter path to other examiners in the network, which may allow them to quickly seek assistance or exchange information that can help them make decisions more efficiently. It is also important to keep in mind that the adjusted R-squared value of this model is very small, indicating that other factors likely have a stronger influence on the time it takes to process applications.

Regression with the Interaction Terms for 'gender x centrality'

```
m2 <- lm(app_proc_time_std ~ examiner_class + gender + race + tenure_days_std +
           degree_std + betweenness_std + closeness_std + degree_std:gender +
           betweenness_std:gender + closeness_std:gender, data = applications2)
summary(m2)
```

```
##
## Call:
## lm(formula = app_proc_time_std ~ examiner_class + gender + race +
##     tenure_days_std + degree_std + betweenness_std + closeness_std +
##     degree_std:gender + betweenness_std:gender + closeness_std:gender,
##     data = applications2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6572 -0.6650 -0.1672  0.4616  7.9123
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -8.607e-02  3.112e-03 -27.657  < 2e-16 ***
## examiner_class17           -4.713e-02  2.809e-03 -16.777  < 2e-16 ***
## examiner_class21            3.022e-01  3.271e-03  92.398  < 2e-16 ***
## examiner_class24            4.098e-01  3.678e-03 111.427  < 2e-16 ***
## gendermale                 -4.522e-02  2.317e-03 -19.515  < 2e-16 ***
## raceblack                  -7.252e-02  5.813e-03 -12.476  < 2e-16 ***
## raceHispanic                5.815e-02  7.400e-03   7.858 3.90e-15 ***
## raceother                   2.965e-01  3.046e-02   9.734  < 2e-16 ***
## racewhite                   1.122e-02  2.439e-03   4.601 4.21e-06 ***
## tenure_days_std            -2.690e-02  1.071e-03 -25.118  < 2e-16 ***
## degree_std                  3.455e-02  2.200e-03  15.701  < 2e-16 ***
## betweenness_std             3.488e-05  2.204e-03   0.016   0.9874
## closeness_std               4.452e-03  1.890e-03   2.355   0.0185 *
## gendermale:degree_std      -2.888e-02  2.496e-03 -11.570  < 2e-16 ***
## gendermale:betweenness_std  3.579e-02  2.498e-03  14.325  < 2e-16 ***
## gendermale:closeness_std   -9.974e-03  2.268e-03  -4.398 1.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9809 on 906598 degrees of freedom
## Multiple R-squared:  0.03787,    Adjusted R-squared:  0.03786
## F-statistic:  2379 on 15 and 906598 DF,  p-value: < 2.2e-16
```

The regression output after including the interaction term 'gender x centrality':

- does not change much for the correlation between the application processing time and the dependent variables for class, gender, race, and tenure days.
- changes for the closeness centrality which is now positively correlated. The interaction terms 'male x degree' & 'male x closeness' are negatively correlated and 'male x betweenness' is positively correlated. This change in the relationship between closeness centrality and application processing time suggests that the relationship between these two variables may be moderated by the gender of the examiner. In particular, the positive correlation between closeness centrality and application processing time may be

driven by female examiners, while male examiners still exhibit a negative correlation. Similarly, male examiners also exhibit a negative correlation between degree centrality and the application processing time, while female examiners do not. However, once again the adjusted R-squared value of the model is very small, indicating that other factors likely have a stronger influence on the time it takes to process applications.