ex4-ona

Jessica Quansah

2024-04-09

Load the data

The dataset will be loaded fromt he cleaned dataset from exercise 3 which would have the gender, race and tenure days already added

```
data_path = "~/GitHub/desktop-tutorial/Exercise-3/"
edges <- read_csv(paste0(data_path,"edges.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.
```

```
applications <- read_csv(paste0(data_path,"cleaned_applications.csv"))
```

```
## Rows: 2018477 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (11): application_number, examiner_name_last, examiner_name_first, exam...
## dbl   (5): examiner_id, examiner_art_unit, appl_status_code, tc, tenure_days
## date  (5): filing_date, patent_issue_date, abandon_date, earliest_date, late...
##
## 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.
```

Creating the application Processing Time

We will now create a variable that shows the processing times of each application. To do this, I first take out all the pending applications because those have no date in either filing or issue date. Then I create a decision date which pulls the patent issue date if the patent was issued or abandon date if the patent was abadoned . The application processing date is created by finding the difference between the filing date and the decision date.

I will then subset the data collecting only the variables I feel will be necessary for the model moving forward and for that I decided on:

-gender -race -tenure days -application processing time disposal type

At the end I remove any missing values

```
applications <- applications[applications$disposal_type != "PEN", ]
app_subset <- subset(applications, disposal_type %in% c("ISS", "ABN"))

#create application processing time which if disposal type is iss (filing - issue date), and abn (filin
app_subset$decision_date <- ifelse(app_subset$disposal_type == "ISS", app_subset$patent_issue_date,
                             ifelse(app_subset$disposal_type == "ABN", app_subset$abandon_date, NA))

app_subset$app_proc_time <- as.numeric(difftime(as.Date(app_subset$decision_date),
                                                 as.Date(app_subset$filing_date),
                                                 units = "days"))

app_subset<-app_subset%>%
  select(application_number, examiner_id, gender, race, tenure_days, app_proc_time, disposal_type)

# Count NA values per column
count_missing <- colSums(is.na(app_subset))
print(count_missing)
```

```
## application_number        examiner_id             gender               race
##                  0               3754             253884                  0
##        tenure_days      app_proc_time      disposal_type
##              18250                  7                  0
```

```
app_subset<-drop_na(app_subset)
```

###Other Checks From here, I just wanted to take a look at some statistics to check the data. I noticed that there were some negative values in the dataset so I dropped those. - Note: I counted the number of missing values and it was only 35 rows so it should not be a problem

```
#How many have a negative value - Why would I have a negative value
summaryd<-summary(app_subset)
print(summaryd)
```

```
##  application_number  examiner_id       gender              race
##  Length:1422804     Min.   :59012   Length:1422804     Length:1422804
##  Class :character   1st Qu.:66605   Class :character   Class :character
##  Mode  :character   Median :75367   Mode  :character   Mode  :character
##                     Mean   :78858
##                     3rd Qu.:93823
##                     Max.   :99988
##   tenure_days   app_proc_time    disposal_type
##  Min.   : 216   Min.   :-13636   Length:1422804
##  1st Qu.:5137   1st Qu.:   768   Class :character
##  Median :6189   Median :  1078   Mode  :character
##  Mean   :5643   Mean   :  1190
##  3rd Qu.:6338   3rd Qu.:  1477
##  Max.   :6518   Max.   :  6255
```

```r
#Drop negative values from app_proc_time
app_subset <- app_subset[app_subset$app_proc_time >= 0, ]
```

###Creating Edges Dataset and computing centrality measures Here edges dataset is created based on applcaitions that are still in the original dataset. Betweenness, Closeness and Degree Centrality are also computed and uploaded into a centrality dataframe. This should enable us to merge with the application dataset in the next step

```r
#Create edges dataset
edges<- edges %>%
  filter(application_number %in% app_subset$application_number)
colnames(edges)[3:4] <- c("from", "to")
#edges<-edges%>%
#  select(from, to)
edges<-na.omit(edges)  #Remove NA values
edges <- edges[, c("from", "to", setdiff(names(edges), c("from", "to")))]

g <- graph_from_data_frame(edges, directed = TRUE)
degree <- degree(g)                           # Degree centrality
closeness <- closeness(g)                     # Closeness centrality
betweenness <- betweenness(g)                 # Betweeness Centrality

#Creation of
centrality <- data.frame(examiner_id    = V(g)$name,
                         degree      = degree,
                         closeness   = closeness,
                         betweenness = betweenness)
```

###Merging and Cleaning The centrality datasets created before were then merged with my application dataset created from before on examiner id. After I checked for any missing values. I also created dummy variables for the model that will be created in the next step

```r
#Merge centralities with app_subset
merged_data <- merge(app_subset, centrality, by= "examiner_id")


# Count NA values per column
count_missing <- colSums(is.na(merged_data))
print(count_missing)
```

```
##        examiner_id application_number             gender               race
##                  0                  0                  0                  0
##        tenure_days      app_proc_time      disposal_type             degree
##                  0                  0                  0                  0
##          closeness        betweenness
##             307878                  0
```

```r
#Regression
#Dummy variables
merged_data$gender= as.factor(merged_data$gender)
merged_data$race= as.factor(merged_data$race)

merged_data<-drop_na(merged_data)
```

###Building Model Here I first start with a model with all the centrality measures and tenure days and disposal type.

```r
reg1 <- lm(app_proc_time ~ gender+betweenness+closeness+degree+tenure_days+disposal_type, data=merged_da
summary(reg1)
```

```
##
## Call:
## lm(formula = app_proc_time ~ gender + betweenness + closeness +
##     degree + tenure_days + disposal_type, data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1447.3  -440.2  -119.3   305.2  4972.0
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.509e+03  8.021e+00  188.08   <2e-16 ***
## gendermale         1.907e+01  1.833e+00   10.40   <2e-16 ***
## betweenness        3.244e-03  1.421e-04   22.84   <2e-16 ***
## closeness         -1.253e+02  2.447e+00  -51.23   <2e-16 ***
## degree            -4.355e-01  2.544e-02  -17.12   <2e-16 ***
## tenure_days       -4.151e-02  1.343e-03  -30.91   <2e-16 ***
## disposal_typeISS  2.692e+01  1.769e+00   15.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 645.7 on 583719 degrees of freedom
## Multiple R-squared:  0.009493,   Adjusted R-squared:  0.009482
## F-statistic: 932.3 on 6 and 583719 DF,  p-value: < 2.2e-16
```

The output of this tells us there is a positive relationship between betweenness centrality and how long one's application takes. It is however a small positive relationship but statistically. There is however a negative relationship between closeness and degree centrality and the application processing time. The R-squared for this model is however really low meaning that it most likely does not explain the variation in the data. I tried carious variations of this but this was the highest. Despite all of them being almost 0.

###Including Interaction Variable The next step was to include the interaction variable on gender so this was done iteratively for each centrality measure. The results however does not change much either on the R-squared or on the relationship between the variables. INterestingly when the interaction is included on all the centrality measureness betweenness centrality is no longer a significant relationship

```r
reg2 <- lm(app_proc_time ~ gender*betweenness+closeness+degree+tenure_days+disposal_type, data=merged_da
summary(reg2)
```

```
##
## Call:
## lm(formula = app_proc_time ~ gender * betweenness + closeness +
##     degree + tenure_days + disposal_type, data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1536.7  -440.3  -119.3   305.4  4969.6
```

```
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.511e+03  8.022e+00 188.336   <2e-16 ***
## gendermale             1.569e+01  1.852e+00   8.475   <2e-16 ***
## betweenness           -6.087e-05  2.975e-04  -0.205    0.838
## closeness             -1.258e+02  2.447e+00 -51.404   <2e-16 ***
## degree                -4.434e-01  2.544e-02 -17.429   <2e-16 ***
## tenure_days           -4.149e-02  1.342e-03 -30.905   <2e-16 ***
## disposal_typeISS       2.741e+01  1.769e+00  15.497   <2e-16 ***
## gendermale:betweenness  4.232e-03  3.348e-04  12.642   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 645.6 on 583718 degrees of freedom
## Multiple R-squared:  0.009764,   Adjusted R-squared:  0.009752
## F-statistic: 822.2 on 7 and 583718 DF,  p-value: < 2.2e-16
```

```
reg3 <- lm(app_proc_time ~ betweenness+gender*closeness+degree+tenure_days+disposal_type, data=merged_da
summary(reg3)
```

```
##
## Call:
## lm(formula = app_proc_time ~ betweenness + gender * closeness +
##     degree + tenure_days + disposal_type, data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1447.9  -440.3  -119.3   305.2  4965.3
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.504e+03  8.176e+00 183.951  < 2e-16 ***
## betweenness          3.233e-03  1.421e-04  22.752  < 2e-16 ***
## gendermale           2.380e+01  2.446e+00   9.732  < 2e-16 ***
## closeness           -1.155e+02  4.161e+00 -27.760  < 2e-16 ***
## degree              -4.350e-01  2.544e-02 -17.099  < 2e-16 ***
## tenure_days         -4.129e-02  1.345e-03 -30.707  < 2e-16 ***
## disposal_typeISS     2.706e+01  1.769e+00  15.293  < 2e-16 ***
## gendermale:closeness -1.448e+01  4.955e+00  -2.923  0.00346 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 645.7 on 583718 degrees of freedom
## Multiple R-squared:  0.009507,   Adjusted R-squared:  0.009495
## F-statistic: 800.4 on 7 and 583718 DF,  p-value: < 2.2e-16
```

```
reg4 <- lm(app_proc_time ~ betweenness+closeness+gender*degree+tenure_days+disposal_type, data=merged_da
summary(reg3)
```

```
##
## Call:
## lm(formula = app_proc_time ~ betweenness + gender * closeness +
```

```
##     degree + tenure_days + disposal_type, data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1447.9  -440.3  -119.3   305.2  4965.3
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.504e+03  8.176e+00 183.951  < 2e-16 ***
## betweenness             3.233e-03  1.421e-04  22.752  < 2e-16 ***
## gendermale              2.380e+01  2.446e+00   9.732  < 2e-16 ***
## closeness              -1.155e+02  4.161e+00 -27.760  < 2e-16 ***
## degree                 -4.350e-01  2.544e-02 -17.099  < 2e-16 ***
## tenure_days            -4.129e-02  1.345e-03 -30.707  < 2e-16 ***
## disposal_typeISS        2.706e+01  1.769e+00  15.293  < 2e-16 ***
## gendermale:closeness -1.448e+01  4.955e+00  -2.923  0.00346 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 645.7 on 583718 degrees of freedom
## Multiple R-squared:  0.009507,    Adjusted R-squared:  0.009495
## F-statistic: 800.4 on 7 and 583718 DF,  p-value: < 2.2e-16
```

```
reg5 <- lm(app_proc_time ~ gender*betweenness + gender*closeness + gender*degree +tenure_days, data=merg
summary(reg5)
```

```
##
## Call:
## lm(formula = app_proc_time ~ gender * betweenness + gender *
##     closeness + gender * degree + tenure_days, data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1525.4  -440.6  -119.5   306.1  4978.5
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.494e+03  8.319e+00 179.641  < 2e-16 ***
## gendermale                4.211e+01  2.977e+00  14.144  < 2e-16 ***
## betweenness              -4.453e-04  3.010e-04  -1.479    0.139
## closeness                -1.061e+02  4.340e+00 -24.459  < 2e-16 ***
## degree                    2.358e-01  5.436e-02   4.338 1.44e-05 ***
## tenure_days              -3.917e-02  1.342e-03 -29.179  < 2e-16 ***
## gendermale:betweenness  4.768e-03  3.410e-04  13.984  < 2e-16 ***
## gendermale:closeness   -2.547e+01  5.203e+00  -4.895 9.82e-07 ***
## gendermale:degree      -8.608e-01  6.154e-02 -13.988  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 645.6 on 583717 degrees of freedom
## Multiple R-squared:  0.00969,    Adjusted R-squared:  0.009676
## F-statistic: 713.9 on 8 and 583717 DF,  p-value: < 2.2e-16
```

Interpreting Results

The model suggest significant relationships between the centrality measures and the application processing times. Intutively these make sense. Higher degree centrality implies you are more connected in the network and as such that could mean you can easily reach the right people to debug any issues you may have while reviewing the applcation.

The coefficient associated with betweenness centrality indicates the effect of a reviewer's position in the advisory network on the processing time of patent applications. A higher betweenness centrality suggests that the reviewer serves as a bridge between other reviewers in the network. The positive coefficient implies that applications reviewed by individuals with higher betweenness centrality tend to have longer processing times. This could be because these reviewers are involved in a larger number of advisory interactions, which might lead to delays in decision-making or coordination issues.

Degree: The coefficient associated with degree centrality reflects the effect of a reviewer's overall connectivity in the advisory network on the processing time of patent applications. A higher degree centrality indicates that the reviewer is connected to more other reviewers in the network. The negative coefficient suggests that applications reviewed by reviewers with higher connectivity tend to have shorter processing times. This could be because well-connected reviewers might have better access to resources, information, or expertise, leading to more efficient reviews.

Closeness: The coefficient associated with closeness centrality indicates the effect of a reviewer's average distance to other reviewers in the advisory network on the processing time of patent applications. A higher closeness centrality suggests that the reviewer is closer to other reviewers in terms of advisory interactions. The negative coefficient implies that applications reviewed by reviewers with higher closeness centrality tend to have shorter processing times. This could be because closer proximity facilitates faster communication, coordination, and decision-making among reviewers.

Gender: The coefficient associated with gender indicates the effect of reviewer gender on the processing time of patent applications. The coefficient for gendermale suggests that male reviewers tend to have longer processing times compared to female reviewers. However, it's important to interpret this result cautiously and consider potential confounding factors or biases in the review process.

Tenure Days: The coefficient associated with tenure days reflects the effect of reviewer tenure (length of service) on the processing time of patent applications. The negative coefficient implies that longer tenure is associated with shorter processing times. This could be because more experienced reviewers might have better knowledge, skills, and efficiency in the review process.

Interaction Term: The coefficient associated with the interaction between gender and closeness (gendermale:closeness) indicates whether the effect of closeness centrality on processing time varies depending on the reviewer's gender. The negative coefficient suggests that the effect of closeness centrality on processing time is attenuated for male reviewers compared to female reviewers. This interaction effect warrants further investigation to understand potential gender-related differences in the impact of network centrality on processing time.