Exercise2

2024-04-08

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
```

import library

import dataset

```
applications <- read_parquet("C:/Users/xzhu71/Desktop/app_data_sample.parquet")
edges <- read_csv(paste0("C:/Users/xzhu71/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.
```

Now let's following what we did in exercise 3, add gender, race and tenure days for examiners.

```
examiner_surnames <- applications %>%
  select(surname = examiner_name_last) %>%
  distinct()
examiner_race <- predict_race(voter.file = examiner_surnames, surname.only = T) %>%
  as_tibble()
```

```
## Predicting race for 2020

## Warning: Unknown or uninitialised column: `state`.

## Proceeding with last name predictions...

## i All local files already up-to-date!

## 701 (18.4%) individuals' last names were not matched.
```

```r
examiner_race <- examiner_race %>%
  mutate(max_race_p = pmax(pred.asi, pred.bla, pred.his, pred.oth, pred.whi)) %>%
  mutate(race = case_when(
    max_race_p == pred.asi ~ "Asian",
    max_race_p == pred.bla ~ "black",
    max_race_p == pred.his ~ "Hispanic",
    max_race_p == pred.oth ~ "other",
    max_race_p == pred.whi ~ "white",
    TRUE ~ NA_character_
  ))
examiner_race <- examiner_race %>%
  select(surname,race)

applications <- applications %>%
  left_join(examiner_race, by = c("examiner_name_last" = "surname"))
```

Question 1

'app_proc_time' variable measures the number of days from application filing data. I plan to check how many categories the disposal_type, and check if the 'PEND' in disposal_type will cause the in both abandon_date and patent_issue_date.

```r
# Examine categories in 'disposal_type'
disposal_type_categories <- table(applications$disposal_type)
disposal_type_categories
```

```
##
##     ABN     ISS    PEND
##  601411 1087306  329760
```

```r
# Checking if 'PEND' in 'disposal_type' corresponds to NA in both 'patent_issue_date' and 'abandon_date
pending_analysis <- applications %>%
  filter(disposal_type == "PEND") %>%
  summarize(
    Count_NA_both = sum(is.na(patent_issue_date) & is.na(abandon_date))
  )
```

The 'pending_analysis' gives 329760 records of both patent_issue data and abandon_date by showing 'pending'.

now calculate the processing time which measures by patent_issue_date/abandon_date substract the filling data

create 'app_proc_time' variable

```r
# Calculate processing time
applications <- applications %>%
  mutate(final_decision_date = coalesce(patent_issue_date, abandon_date),
         app_proc_time = as.integer(final_decision_date - filing_date))
```

deal with missing values

now check the number of 'app_proc_time' NA reocrds

```
na_app_proc_time_count <- sum(is.na(applications$app_proc_time))
na_app_proc_time_count
```

```
## [1] 329761
```

There are 329761 'NA' records, but the data is not consistent than we calcualte before, it has 1 difference. To investigate the discrepancy further and identify the specific records contributing to the difference in counts,

```
# First, add a column to indicate if both important dates are NA
data <- applications %>%
  mutate(BothDatesNA = is.na(patent_issue_date) & is.na(abandon_date))

# Now, let's find records that contribute to the NA in 'app_proc_time' but are not 'PEND'
non_pend_na_proc_time <-applications %>%
  filter(is.na(app_proc_time) & disposal_type != "PEND")

# Additionally, let's explore cases where 'app_proc_time' is NA to understand all possible reasons
na_proc_time_exploration <- applications %>%
  filter(is.na(app_proc_time))

print("Non-PEND with NA app_proc_time:")
```

```
## [1] "Non-PEND with NA app_proc_time:"
```

```
print(non_pend_na_proc_time)
```

```
## # A tibble: 1 x 20
##   application_number filing_date examiner_name_last examiner_name_first
##   <chr>              <date>      <chr>              <chr>
## 1 14120992           2014-09-26  FONTAINHAS         AURORA
## # i 16 more variables: examiner_name_middle <chr>, examiner_id <dbl>,
## #   examiner_art_unit <dbl>, uspc_class <chr>, uspc_subclass <chr>,
## #   patent_number <chr>, patent_issue_date <date>, abandon_date <date>,
## #   disposal_type <chr>, appl_status_code <dbl>, appl_status_date <chr>,
## #   tc <dbl>, gender <chr>, race <chr>, final_decision_date <date>,
## #   app_proc_time <int>
```

There is 1 record that shows 'ISS'disposal_type, but the 'patent_issue_date' and 'abadon_date' is non, that is why the 'na' in 'app_proc_time' has one more record than the number of 'na' in original dataset.

now let's drop missing records in 'app_proc_time' to get clean data set to prepare for the centrality

```
cleaned_data <- applications %>%
  filter(!is.na(app_proc_time))
```

Question 2

create the network

```
# Identify all unique examiners
all_examiners <- unique(c(edges$ego_examiner_id, edges$alter_examiner_id))
#Create the network
g = graph_from_data_frame(edges[, c("ego_examiner_id", "alter_examiner_id")],
                          directed = TRUE,
                          vertices = data.frame(name = all_examiners ))
```

```
## Warning in graph_from_data_frame(edges[, c("ego_examiner_id",
## "alter_examiner_id")], : In 'd' 'NA' elements were replaced with string "NA"
```

```
## Warning in graph_from_data_frame(edges[, c("ego_examiner_id",
## "alter_examiner_id")], : In 'vertices[,1]' 'NA' elements were replaced with
## string "NA"
```

now i want to calculate the 'in' and 'out' centrality seperately

```
# Calculate degree centrality (both in-degree and out-degree)
edges_processed <- edges %>%
  group_by(ego = ego_examiner_id, alter = alter_examiner_id) %>%
  summarise(application_count = n_distinct(application_number), .groups = "drop")
g <- graph_from_data_frame(d = edges_processed, directed = TRUE)
```

```
## Warning in graph_from_data_frame(d = edges_processed, directed = TRUE): In 'd'
## 'NA' elements were replaced with string "NA"
```

```
in_degree_centrality <- degree(g, mode = "in")
out_degree_centrality <- degree(g, mode = "out")


betweenness_centrality <- betweenness(g, directed = TRUE)


in_closeness_centrality <- closeness(g, mode = "in", weights = E(g)$application_count)
out_closeness_centrality <- closeness(g, mode = "out", weights = E(g)$application_count)


centrality_measures <- data.frame(
  examiner_id = V(g)$name,
  in_degree = in_degree_centrality,
  out_degree = out_degree_centrality,
  betweenness = betweenness_centrality,
  in_closeness = in_closeness_centrality,
  out_closeness = out_closeness_centrality
)
```

After buliding up the centrality, let's try different combination of lm model

```
cleaned_data <- cleaned_data %>%
  mutate(examiner_id = as.character(examiner_id))
joined_data <- cleaned_data %>%
  left_join(centrality_measures, by = "examiner_id")
```

```r
lm_model <- lm(app_proc_time ~ in_degree + out_degree + betweenness + in_closeness + out_closeness, data

summary(lm_model)
```

```
##
## Call:
## lm(formula = app_proc_time ~ in_degree + out_degree + betweenness +
##     in_closeness + out_closeness, data = joined_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2526.8  -428.6  -110.3   294.4  4578.9
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.221e+03  2.136e+00 571.411  < 2e-16 ***
## in_degree      2.542e+00  2.411e-01  10.542  < 2e-16 ***
## out_degree     1.978e+00  1.510e-01  13.097  < 2e-16 ***
## betweenness    5.580e-04  1.617e-04   3.450 0.000561 ***
## in_closeness   2.522e+01  3.102e+00   8.129 4.34e-16 ***
## out_closeness -1.113e+02  3.069e+00 -36.263  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 626.5 on 364474 degrees of freedom
##   (1324236 observations deleted due to missingness)
## Multiple R-squared:  0.007623,   Adjusted R-squared:  0.007609
## F-statistic: 559.9 on 5 and 364474 DF,  p-value: < 2.2e-16
```

Interpretation: Intercept (1.221e+03): The model intercept is approximately 1221. This value represents the expected application processing time without any other elemeents involved is 1221.

Centrality Measures In-Degree (2.542e+00): The coefficient for in_degree is 2.542, indicating that for each additionalincoming connection, the estimated application processing time increases by approximately 2.542 units. This suggests that examiners with more incoming connections have slightly longer processing times and this elements play important role affecting the processing time.

Out-Degree (1.978e+00): The out_degree coefficient is 1.978, showing that for each additional outgoing connection, the estimated average processing time increases by about 1.978 units.

Betweenness (5.580e-04): The betweenness centrality has a coefficient of 0.000558, suggesting a minor increase in app_proc_time with higher betweenness centrality. Since the effect size is small, that might nit be an efficient indicators.

In-Closeness (2.522e+01): The processing time is observed to increase by 25.22 units with each unit increase in in-closeness centrality. This could reflect the burden of being frequently consulted or involved in many processes.

Out-Closeness (-1.113e+02): The coefficient for out_closeness centrality is -111.3, indicating a significant decrease in app_proc_time with increased out-closeness centrality. This might imply that examiners who can reach out to the network more efficiently tend to have shorter processing times, possibly due to better access to information or resources.

```r
# Constructing the linear regression model
lm_model2 <- lm(app_proc_time ~ in_degree + out_degree + betweenness + in_closeness + out_closeness + e
```

```r
# Viewing the summary of the regression model
summary(lm_model2)
```

```
##
## Call:
## lm(formula = app_proc_time ~ in_degree + out_degree + betweenness +
##     in_closeness + out_closeness + examiner_art_unit + gender +
##     race, data = joined_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2358.2  -408.9   -97.4   286.8  4385.7
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.486e+02  8.497e+00  41.024  < 2e-16 ***
## in_degree         -2.159e+00  2.508e-01  -8.607  < 2e-16 ***
## out_degree         5.369e-01  1.525e-01   3.520 0.000432 ***
## betweenness        1.526e-03  1.625e-04   9.393  < 2e-16 ***
## in_closeness       2.883e+01  3.286e+00   8.773  < 2e-16 ***
## out_closeness     -7.715e+01  3.243e+00 -23.793  < 2e-16 ***
## examiner_art_unit  4.660e-01  4.127e-03 112.924  < 2e-16 ***
## gendermale         6.405e+00  2.422e+00   2.644 0.008188 **
## raceblack         -1.901e+01  7.146e+00  -2.660 0.007808 **
## raceHispanic       1.320e+02  1.072e+01  12.312  < 2e-16 ***
## racewhite         -3.182e+01  2.521e+00 -12.621  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 608.6 on 317691 degrees of freedom
##   (1371014 observations deleted due to missingness)
## Multiple R-squared:  0.05593,    Adjusted R-squared:  0.0559
## F-statistic:  1882 on 10 and 317691 DF,  p-value: < 2.2e-16
```

Additional Variables Impact by comparing with the model 1 Examiner Art Unit (0.466): The positive coefficient for examiner_art_unit by the significant p-value suggests with each unit increase in art unit number associated with a slight increase in processing time. This may reflect differences in application complexity or volume across different art units.

Gender Male (6.405): The inclusion of gender with a positive coefficient for males indicates that gender differences exist in processing times, with male examiners having longer processing times compared to their counterparts.

Race Effects: The coefficients for race reveal significant differences in processing times across racial groups, with Hispanic examiners (132) experiencing substantially longer processing times, Black(-19.01) and white (-31.82 units) examiners shorter ones compared to the baseline group. This suggests that racial demographics influence processing times, potentially due to differences in experience, workload distribution, or other systemic factors within the USPTO.

Question 3 & 4

```r
lm_model3 <- lm(app_proc_time ~ in_degree * gender + out_degree * gender + betweenness * gender + in_cl

summary(lm_model3)
```

```
##
## Call:
## lm(formula = app_proc_time ~ in_degree * gender + out_degree *
##     gender + betweenness * gender + in_closeness * gender + out_closeness *
##     gender + examiner_art_unit + race, data = joined_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2352.9  -408.5   -97.1   286.7  4391.7
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.430e+02  9.075e+00  37.793  < 2e-16 ***
## in_degree                   1.518e+00  4.919e-01   3.087 0.002025 **
## gendermale                  1.875e+01  4.896e+00   3.830 0.000128 ***
## out_degree                  1.883e+00  3.303e-01   5.701 1.19e-08 ***
## betweenness                -1.924e-03  3.664e-04  -5.251 1.52e-07 ***
## in_closeness                1.657e+01  5.557e+00   2.982 0.002866 **
## out_closeness              -6.703e+01  6.001e+00 -11.169  < 2e-16 ***
## examiner_art_unit           4.628e-01  4.139e-03 111.811  < 2e-16 ***
## raceblack                  -1.781e+01  7.212e+00  -2.469 0.013545 *
## raceHispanic                1.353e+02  1.075e+01  12.581  < 2e-16 ***
## racewhite                  -3.095e+01  2.528e+00 -12.242  < 2e-16 ***
## in_degree:gendermale       -4.482e+00  5.722e-01  -7.833 4.76e-15 ***
## gendermale:out_degree      -1.489e+00  3.722e-01  -4.000 6.33e-05 ***
## gendermale:betweenness      4.265e-03  4.089e-04  10.430  < 2e-16 ***
## gendermale:in_closeness     1.629e+01  6.857e+00   2.376 0.017517 *
## gendermale:out_closeness   -1.068e+01  7.092e+00  -1.506 0.131968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 608.5 on 317686 degrees of freedom
##   (1371014 observations deleted due to missingness)
## Multiple R-squared:  0.0563, Adjusted R-squared:  0.05625
## F-statistic:  1263 on 15 and 317686 DF,  p-value: < 2.2e-16
```

Key findings: The interaction terms (in_degree:gendermale, gendermale:out_degree, gendermale:betweenness, gendermale:in_closeness) significantly modify the relationship between centrality measures and app_proc_time.

in_degree: A base increase of 1.518 in app_proc_time for each additional in-degree point, suggesting examiners with more incoming connections experience slightly longer processing times. gendermale: Male examiners have a increase of 18.75 in processing times compared to female examiners, indicating a gender-based disparity in processing times. out_degree, betweenness, in_closeness, and out_closeness show expected effects on app_proc_time, similar to previous models, indicating their individual impacts on processing times.

(in_degree:gendermale): The negative coefficient (-4.482) implies that the effect of in-degree on processing time is reduced for male examiners compared to female examiners. This indicated that while in-degree generally increases processing times, the increase is less significant for males.

(gendermale:out_degree): The negative coefficient (-1.489) for this interaction term suggests that the positive effect of out-degree on processing time is also less for male examiners.

(gendermale:betweenness): The positive interaction term (4.265e-03) indicates that the minor increase in processing time associated with betweenness centrality is more significant for male examiners.

(gendermale:in_closeness and gendermale:out_closeness): These terms indicate differential effects of closeness centrality on processing times by gender, with in-closeness increasing processing times more for males and the negative effect of out-closeness on processing times being less significant for males.

Implications for the USPTO

Gender Differences:The findings highlight significant gender differences in how network centrality affects processing time. This reflect underlying the differences in work allocation, collaboration patterns, or responsibilities between male and female reviewers. Ensure a balanced gender ratio within the team, especially in those projects or departments that involve a high degree of collaboration and communication. Not only does this facilitate the exchange of different perspectives, but it also helps balance work loads and responsibilities. Encourage and support female employees to take on core roles within the network through training and career development opportunities.