Ex4 - Centrality and efficiency

Jennifer Liu

2024-04-09

Step0: Import Libraries and dataset

```
library(readr)
library(dplyr)
```

```
##
## 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(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(igraph)
```

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

```
## The following objects are masked from 'package:lubridate':
##
##     %--%, union
```

```
## 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(gender)
library(arrow)
```

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

```
## The following object is masked from 'package:lubridate':
##
##     duration
```

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

```
library(dplyr)
library(wru)
```

```
##
## Please cite as:
##
## Khanna K, Bertelsen B, Olivella S, Rosenman E, Rossell Hayes A, Imai K
## (2024). _wru: Who are You? Bayesian Prediction of Racial Category Using
## Surname, First Name, Middle Name, and Geolocation_. R package version
## 3.0.1, <https://CRAN.R-project.org/package=wru>.
##
## Note that wru 2.0.0 uses 2020 census data by default.
## Use the argument `year = "2010"`, to replicate analyses produced with earlier package version
s.
```

```
library(ggplot2)
library(ggraph)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ forcats 1.0.0     ✓ tibble  3.2.1
## ✓ purrr   1.0.2     ✓ tidyr   1.3.1
## ✓ stringr 1.5.1
```

```
## ── Conflicts ─────────────────────────────────────── tidyverse_conflicts() ──
## ✖ igraph::%--%()          masks lubridate::%--%()
## ✖ tibble::as_data_frame() masks igraph::as_data_frame(), dplyr::as_data_frame()
## ✖ purrr::compose()        masks igraph::compose()
## ✖ tidyr::crossing()       masks igraph::crossing()
## ✖ arrow::duration()       masks lubridate::duration()
## ✖ dplyr::filter()         masks stats::filter()
## ✖ dplyr::lag()            masks stats::lag()
## ✖ purrr::simplify()       masks igraph::simplify()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(tidygraph)
```

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

```
applications = read_parquet("C:\\Users\\Admin\\Downloads\\672_project_data\\app_data_sample.parq
uet")

#review first rows of app_data_sample parquet file
head(applications)
```

```
## # A tibble: 6 × 16
##   application_number filing_date examiner_name_last examiner_name_first
##   <chr>              <date>      <chr>              <chr>
## 1 08284457           2000-01-26  HOWARD             JACQUELINE
## 2 08413193           2000-10-11  YILDIRIM           BEKIR
## 3 08531853           2000-05-17  HAMILTON           CYNTHIA
## 4 08637752           2001-07-20  MOSHER             MARY
## 5 08682726           2000-04-10  BARR               MICHAEL
## 6 08687412           2000-04-28  GRAY               LINDA
## # ℹ 12 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>
```

```
edges = read.csv("C:\\Users\\Admin\\Downloads\\672_project_data\\edges_sample.csv")

edges <- edges %>%
  mutate(
    from = as.character(ego_examiner_id), # Convert IDs to character
    to = as.character(alter_examiner_id)
  ) %>%
  drop_na()

#review first rows of edges_sample csv file
head(edges)
```

```
##   application_number advice_date ego_examiner_id alter_examiner_id  from    to
## 1            9402488  2008-11-17           84356             66266 84356 66266
## 2            9402488  2008-11-17           84356             63519 84356 63519
## 3            9402488  2008-11-17           84356             98531 84356 98531
## 4            9445135  2008-08-21           92953             71313 92953 71313
## 5            9445135  2008-08-21           92953             93865 92953 93865
## 6            9445135  2008-08-21           92953             91818 92953 91818
```

Introduce gender, race, tenure variables

```
examiner_names = applications %>% distinct(examiner_name_first)

examiner_names_gender <- examiner_names %>%
  do(results = gender(.$examiner_name_first, method = "ssa")) %>%
  unnest(cols = c(results), keep_empty = TRUE) %>%
  select(
    examiner_name_first = name,
    gender,
    proportion_female
  ) %>%
  filter(!is.na(gender))

examiner_names_gender <- examiner_names_gender %>%
  select(examiner_name_first, gender)

applications <- applications %>%
  left_join(examiner_names_gender, by = "examiner_name_first")

# cleaning up
rm(examiner_names)
rm(examiner_names_gender)
gc()
```

```
##            used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells  4731970 252.8    8279360 442.2  4985657 266.3
## Vcells 49966614 381.3   95971004 732.3 80282618 612.6
```

```
examiner_surnames <- applications %>%
  select(surname = examiner_name_last) %>%
  distinct()

examiner_race <- examiner_surnames %>%
  filter(!is.na(surname)) %>%
  predict_race(voter.file = ., surname.only = TRUE) %>%
  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.
```

```
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"))

#Cleaning up
rm(examiner_race)
rm(examiner_surnames)
gc()
```

```
##            used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells  4808004 256.8    8279360 442.2  6179523 330.1
## Vcells 52136721 397.8   95971004 732.3 95176987 726.2
```

```
examiner_dates <- applications %>%
  select(examiner_id, filing_date, appl_status_date)

examiner_dates <- examiner_dates %>%
  mutate(start_date = ymd(filing_date), end_date = as_date(dmy_hms(appl_status_date)))

examiner_tenure <- examiner_dates %>%
  # Remove rows with NA in start_date or end_date before grouping and summarising
  filter(!is.na(start_date) & !is.na(end_date)) %>%
  group_by(examiner_id) %>%
  summarise(
    earliest_date = min(start_date, na.rm = TRUE),
    latest_date = max(end_date, na.rm = TRUE),
    tenure_days = interval(earliest_date, latest_date) %/% days(1),
    .groups = 'drop' # Automatically drop the grouping
  ) %>%
  # Keep records with a latest_date before 2018
  filter(year(latest_date) < 2018)

applications <- applications %>%
  left_join(examiner_tenure, by = "examiner_id")

rm(examiner_tenure)
gc()
```

```
##            used  (Mb) gc trigger   (Mb)  max used  (Mb)
## Ncells  4816560 257.3    8279360  442.2   8279360 442.2
## Vcells 68304431 521.2  138374244 1055.8 114619942 874.5
```

Introduce Processing Time

An essential part of our analysis involves determining the application processing time, which is the period between the filing date and the final decision. This measurement is critical for assessing the efficiency of the patent examination process.

First of all, in this analysis, I will exclude applications with “PEND” status.

```
applications <- applications %>%
  filter(disposal_type != "PEND")

applications <- applications %>%
  mutate(app_proc_time = interval(
    ymd(filing_date),
    dmy_hms(appl_status_date)
  ) %/% days(1))
gc()
```

```
##            used  (Mb) gc trigger   (Mb)  max used   (Mb)
## Ncells  4486517 239.7    8279360  442.2   8279360  442.2
## Vcells 62407576 476.2  166129092 1267.5 165401516 1262.0
```

```
head(applications)
```

```
## # A tibble: 6 × 22
##   application_number filing_date examiner_name_last examiner_name_first
##   <chr>              <date>      <chr>              <chr>
## 1 08284457           2000-01-26  HOWARD             JACQUELINE
## 2 08413193           2000-10-11  YILDIRIM           BEKIR
## 3 08531853           2000-05-17  HAMILTON           CYNTHIA
## 4 08637752           2001-07-20  MOSHER             MARY
## 5 08682726           2000-04-10  BARR               MICHAEL
## 6 08687412           2000-04-28  GRAY               LINDA
## # ℹ 18 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>, earliest_date <date>,
## #   latest_date <date>, tenure_days <dbl>, app_proc_time <dbl>
```

Generate Network Graph

Next, we ready the applications dataframe for inclusion in the network graph. This process involves moving the examiner_id to a more accessible location, changing IDs into character strings to match the edge data format, and updating the examiner_id to name for better clarity. We then construct a directed graph using the edges dataframe, integrating examiner information from the applications.

```
# Preparing applications data for graph creation
applications <- applications %>%
  relocate(examiner_id, .before = application_number) %>%
  mutate(examiner_id = as.character(examiner_id)) %>%
  drop_na(examiner_id) %>%
  rename(name = examiner_id)

# Creating a directed graph from the edges data
graph <- tbl_graph(
  edges = (edges %>% relocate(from, to)),
  directed = TRUE
)

# Enriching graph nodes with examiner data from applications
graph <- graph %>%
  activate(nodes) %>%
  inner_join(
    (applications %>% distinct(name, .keep_all = TRUE)),
    by = "name"
  )

# Display the graph structure
graph
```

```
## # A tbl_graph: 2489 nodes and 17720 edges
## #
## # A directed multigraph with 127 components
## #
## # Node Data: 2,489 × 22 (active)
##    name  application_number filing_date examiner_name_last examiner_name_first
##    <chr> <chr>              <date>      <chr>              <chr>
##  1 84356 09402488           2000-02-16  STEADMAN           DAVID
##  2 66266 09509710           2000-06-15  BRUMBACK           BRENDA
##  3 63519 09463947           2000-02-04  WEBER              JON
##  4 98531 09423418           2000-06-22  BRAGDON            KATHLEEN
##  5 92953 09445135           2000-03-13  RAMAN              USHA
##  6 93865 10481715           2004-06-01  WONG               JOSEPH
##  7 91818 09424167           2000-05-30  PILLAI             NAMITHA
##  8 66805 09486723           2000-05-18  PICH               PONNOREAY
##  9 70919 09703038           2000-10-31  SAM                PHIRIN
## 10 72253 09242244           2000-02-29  WOITACH            JOSEPH
## # ℹ 2,479 more rows
## # ℹ 17 more variables: examiner_name_middle <chr>, 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>, earliest_date <date>, latest_date <date>, tenure_days <dbl>,
## #   app_proc_time <dbl>
## #
## # Edge Data: 17,720 × 6
##    from    to application_number advice_date ego_examiner_id alter_examiner_id
##   <int> <int>              <int> <chr>                 <int>             <int>
## 1     1     2            9402488 2008-11-17            84356             66266
## 2     1     3            9402488 2008-11-17            84356             63519
## 3     1     4            9402488 2008-11-17            84356             98531
## # ℹ 17,717 more rows
```

After performing these operations, the enriched graph is displayed, showing that it consists of 2,489 nodes and 17,720 edges. This directed multigraph with 127 components indicates a complex network of interactions among USPTO patent examiners. This enriched graph will serve as a foundation for exploring questions related to the length of patent application prosecution, the role of network structure, and the impact of race and ethnicity on these processes. For instance, such patterns could indicate frequent collaborations or consultations on patent applications, underscoring a complex network of professional relationships within the USPTO. The identification of 127 separate components in the network points to a segmented structure, where certain groups of examiners might interact more closely, likely due to shared specializations or organizational divisions. This segmentation could reflect the varied technical areas that the patent applications encompass, suggesting that examiners are naturally grouped by their expertise or the structural organization of their departments.

Apply Linear Regression Models

Calculate centrality measures

```
node_data <- graph %>%
  activate(nodes) %>%
  mutate(
    degree = centrality_degree(),
    betweenness = centrality_betweenness(),
    closeness = centrality_closeness()
  ) %>%
  arrange(-degree) %>%
  as_tibble() %>%
  mutate(tc = as.factor(tc))

node_data
```

```
## # A tibble: 2,489 × 25
##    name  application_number filing_date examiner_name_last examiner_name_first
##    <chr> <chr>              <date>      <chr>              <chr>
##  1 83670 09856864           2001-07-05  LEE                JAE
##  2 97910 09486362           2000-02-28  COUNTS             GARY
##  3 73920 10373614           2003-02-25  HOBBS              LISA
##  4 67226 09483069           2000-01-14  ZHEN               LI
##  5 80730 10345713           2003-01-16  JOY                DAVID
##  6 75615 09943424           2001-08-30  DECKER             CASSANDRA
##  7 62152 10486872           2004-08-12  SIDDIQUEE          MUHAMMAD
##  8 69098 10491238           2004-11-15  VASISTH            VISHAL
##  9 67690 09504184           2000-02-15  MCINTOSH III       TRAVISS
## 10 74061 10480716           2004-07-02  TRAN               THINH
## # ℹ 2,479 more rows
## # ℹ 20 more variables: examiner_name_middle <chr>, 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 <fct>, gender <chr>,
## #   race <chr>, earliest_date <date>, latest_date <date>, tenure_days <dbl>,
## #   app_proc_time <dbl>, degree <dbl>, betweenness <dbl>, closeness <dbl>
```

```
linear_model = lm(app_proc_time ~ degree + betweenness + closeness + tenure_days, data = node_da
ta)

summary(linear_model)
```

```
##
## Call:
## lm(formula = app_proc_time ~ degree + betweenness + closeness +
##     tenure_days, data = node_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2654.0  -855.3   -68.5   739.1  3219.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  632.22229  147.22113   4.294 1.87e-05 ***
## degree         1.26052    1.86035   0.678    0.498
## betweenness   -0.01013    0.01005  -1.007    0.314
## closeness   -102.06212   97.28416  -1.049    0.294
## tenure_days    0.37572    0.02667  14.086  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1228 on 1424 degrees of freedom
##   (1060 observations deleted due to missingness)
## Multiple R-squared:  0.1227, Adjusted R-squared:  0.1202
## F-statistic: 49.79 on 4 and 1424 DF,  p-value: < 2.2e-16
```

Result

The extended linear regression analysis of `app_proc_time` with predictors `degree`, `betweenness`, `closeness`, and `tenure_days` using the `node_data` dataset reveals: - The **intercept** is 632.22, indicating that the base processing time is around 632 days when all other variables are zero, with a significant p-value of 1.87e-05. - The coefficient for `degree` is 1.26, but with a p-value of 0.498, suggesting no significant effect on processing time. - The `betweenness` centrality has a coefficient of -0.01013 with a p-value of 0.314, indicating a non-significant negative relationship with processing time. - The `closeness` centrality shows a coefficient of -102.06, but with a p-value of 0.294, also suggesting a non-significant effect on processing time. - The `tenure_days` variable has a significant positive effect on processing time, with a coefficient of 0.37572 and a very significant p-value (< 2e-16). - The residuals range from -2654 to 3219.3, with a median close to -68.5, pointing to variability in the model's accuracy. - The model has a residual standard error of 1228 and explains about 12.27% of the variance in processing time, as indicated by an R-squared of 0.1227 and an adjusted R-squared of 0.1202. - The F-statistic is 49.79, with a highly significant p-value (< 2.2e-16), suggesting that the overall model, including all four predictors, significantly predicts processing time, although individual predictors vary in their significance and impact.

Add interaction terms

```
linear_model2 = lm(app_proc_time ~ degree + betweenness + closeness + tenure_days + degree*gende
r, data = node_data)

summary(linear_model2)
```

```
##
## Call:
## lm(formula = app_proc_time ~ degree + betweenness + closeness +
##     tenure_days + degree * gender, data = node_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2659.5  -851.6   -70.4   747.0  3314.3
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       531.02235  176.44120   3.010  0.00267 **
## degree              5.29087    3.46326   1.528  0.12685
## betweenness        -0.01013    0.01021  -0.992  0.32150
## closeness         -32.12180  106.31992  -0.302  0.76261
## tenure_days         0.36867    0.02851  12.930  < 2e-16 ***
## gendermale        126.74137   92.80535   1.366  0.17230
## degree:gendermale  -2.93361    4.06588  -0.722  0.47073
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1217 on 1176 degrees of freedom
##   (1306 observations deleted due to missingness)
## Multiple R-squared:  0.1273, Adjusted R-squared:  0.1229
## F-statistic:  28.6 on 6 and 1176 DF,  p-value: < 2.2e-16
```

The regression analysis now incorporates an interaction term between `degree` and `gender`, alongside the other predictors. Key findings include:

- **Base Processing Time (Intercept):** Estimated at 531 days when all other variables are zero, with a significant p-value of 0.00267, indicating a meaningful baseline effect.
- **Degree Effect:** The `degree` coefficient is 5.29, suggesting an increase in processing time per degree unit. However, this is not statistically significant (p-value = 0.12685), hinting at a potential inconsistency in the degree's impact across genders.
- **Betweenness Centrality:** Shows a negligible negative impact on processing time (coefficient = -0.01013) but lacks statistical significance (p-value = 0.32150).
- **Closeness Centrality:** The coefficient of -32.12 indicates a reduction in processing time with higher closeness values, though it is not statistically significant (p-value = 0.76261).
- **Tenure Days Influence:** Demonstrates a strong positive relationship with processing time, with a coefficient of 0.36867 and a highly significant p-value (< 2e-16).
- **Gender (Male) Impact:** Being male is associated with an increase in processing time by 126.74 days, although not statistically significant (p-value = 0.17230).
- **Interaction Term (Degree and Male Gender):** The interaction has a negative coefficient of -2.93361, implying the effect of `degree` on processing time is reduced for males, yet this interaction is not statistically significant (p-value = 0.47073).

Model Fit: - The residual standard error is 1217 days. - The R-squared value is 0.1273, signifying that around 12.73% of the variance in processing time is accounted for by the model. - An F-statistic of 28.6 and a p-value less than 2.2e-16 confirm the statistical significance of the model overall.

In conclusion, the addition of the interaction term slightly improves the explanatory power of the model, but many of the individual effects, including the interaction term itself, do not achieve statistical significance. This outcome suggests that the interaction between `degree` and `gender` does not significantly enhance the understanding of processing time dynamics.

Add interaction terms

```
linear_model3 = lm(app_proc_time ~ degree + betweenness + closeness + tenure_days + gender*betwe
enness, data = node_data)

summary(linear_model3)
```

```
##
## Call:
## lm(formula = app_proc_time ~ degree + betweenness + closeness +
##     tenure_days + gender * betweenness, data = node_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2650.8  -854.8   -78.0   742.8  3276.6
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            548.17388  173.15958   3.166  0.00159 **
## degree                   3.28730    2.00337   1.641  0.10109
## betweenness              0.02176    0.03311   0.657  0.51106
## closeness              -29.66524  106.33914  -0.279  0.78032
## tenure_days              0.36810    0.02851  12.911  < 2e-16 ***
## gendermale             102.83618   79.80989   1.289  0.19782
## betweenness:gendermale  -0.03572    0.03465  -1.031  0.30284
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1216 on 1176 degrees of freedom
##   (1306 observations deleted due to missingness)
## Multiple R-squared:  0.1277, Adjusted R-squared:  0.1233
## F-statistic:  28.7 on 6 and 1176 DF,  p-value: < 2.2e-16
```

In the updated regression analysis of `app_proc_time`, an interaction term between `gender` and `betweenness` has been added to the model, which also includes `degree`, `closeness`, and `tenure_days`. Here are the interpretations:

- **Base Processing Time (Intercept):** The model estimates the baseline processing time at 548.17 days when all predictors are zero, significantly established with a p-value of 0.00159.
- **Degree:** The coefficient for `degree` is 3.29, indicating a slight increase in processing time per unit increase in degree, although this is not statistically significant (p-value = 0.10109).
- **Betweenness:** The coefficient of 0.02176 suggests a marginal increase in processing time with higher betweenness, but this effect is not statistically significant (p-value = 0.51106).
- **Closeness:** The impact of `closeness` is minimal and negative (-29.67) but not significant (p-value = 0.78032), indicating no strong relationship with processing time.

- **Tenure Days:** Shows a strong positive relationship with processing time, as each additional day in tenure increases processing time by 0.3681, highly significant (p-value < 2e-16).
- **Gender (Male):** Being male is associated with an increase in processing time by 102.84 days, but this increase is not statistically significant (p-value = 0.19782).
- **Interaction (Betweenness and Male Gender):** The interaction term has a coefficient of -0.03572, indicating a slight decrease in the effect of betweenness on processing time for males, though this interaction is not statistically significant (p-value = 0.30284).

Model Fit: - The residual standard error is 1216 days. - The R-squared value of 0.1277 suggests that around 12.77% of the variance in processing time is explained by the model, with an adjusted R-squared of 0.1233. - The F-statistic is 28.7 with a highly significant p-value (< 2.2e-16), indicating that the model, as a whole, is statistically significant.

In conclusion, while the model shows a statistically significant overall fit, the effects of individual predictors, including the interaction between `gender` and `betweenness`, are mostly not statistically significant. This suggests that the predictors, as currently specified, have limited individual impact on explaining the variance in processing time.

Add interaction terms

```
linear_model4 = lm(app_proc_time ~ degree + betweenness + closeness + tenure_days + gender*close
ness, data = node_data)

summary(linear_model4)
```

```
##
## Call:
## lm(formula = app_proc_time ~ degree + betweenness + closeness +
##     tenure_days + gender * closeness, data = node_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2651.8  -845.6   -70.2   737.3  3339.7
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           590.49064  180.84478   3.265  0.00113 **
## degree                  3.21516    2.00443   1.604  0.10898
## betweenness            -0.01046    0.01019  -1.027  0.30450
## closeness            -137.20102  190.57398  -0.720  0.47171
## tenure_days             0.36845    0.02852  12.921  < 2e-16 ***
## gendermale             46.93874  104.33274   0.450  0.65287
## closeness:gendermale  144.56922  219.66800   0.658  0.51059
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1217 on 1176 degrees of freedom
##   (1306 observations deleted due to missingness)
## Multiple R-squared:  0.1273, Adjusted R-squared:  0.1228
## F-statistic: 28.58 on 6 and 1176 DF,  p-value: < 2.2e-16
```

An interaction term between `gender` and `closeness` is introduced to the regression analysis of `app_proc_time`, alongside other predictors like `degree`, `betweenness`, and `tenure_days`. The findings are as follows:

- **Base Processing Time (Intercept):** The model estimates the baseline processing time at 590.49 days when other predictors are zero, with a significant p-value of 0.00113.
- **Degree:** The coefficient for `degree` is 3.22, implying a modest increase in processing time per degree unit, but this is not statistically significant (p-value = 0.10898).
- **Betweenness:** The coefficient of -0.01046 for `betweenness` suggests a negligible decrease in processing time, which is not statistically significant (p-value = 0.30450).
- **Closeness:** The coefficient for `closeness` is -137.20, indicating a potential decrease in processing time, but this effect is not significant (p-value = 0.47171).
- **Tenure Days:** A positive significant relationship with processing time is indicated by a coefficient of 0.36845 (p-value < 2e-16), showing increased processing time with longer tenure days.
- **Gender (Male):** Being male is associated with an increase in processing time by 46.94 days, though not significant (p-value = 0.65287).
- **Interaction (Closeness and Male Gender):** The interaction term has a coefficient of 144.57, suggesting that the relationship between `closeness` and processing time might vary by gender. However, this interaction is not statistically significant (p-value = 0.51059).

Model Fit: - The residual standard error of the model is 1217 days. - The R-squared value is 0.1273, and the adjusted R-squared is 0.1228, indicating that about 12.73% of the variance in processing time is explained by the model. - The F-statistic is 28.58, with the model being highly significant overall (p-value < 2.2e-16).

The introduction of the interaction between `gender` and `closeness` has not led to significant changes in the model's explanatory power. While the overall model remains statistically significant, the individual predictors, including the interaction term, show limited significance in explaining the variance in processing time.

Conclusion

Based on the regression analyses, the interaction terms between `gender` and variables such as `closeness` and `betweenness` were not statistically significant. This indicates that the relationship between variables like `degree`, `betweenness`, `closeness`, and `tenure_days` and the application processing time does not substantially differ by examiner gender at the USPTO. Here are the implications:

1. **Gender Neutrality:** The absence of significant gender-related interaction effects suggests that the USPTO's application processing times are generally gender-neutral, reflecting fairness and equality in the examination process.

2. **Importance of Other Factors:** With examiner gender not significantly affecting processing times, the focus should shift to other influencing factors. The significant impact of `tenure_days` on processing time, for example, indicates that experience or service length might be crucial areas to consider for process optimization.

3. **Training and Development:** The positive relationship between `tenure_days` and processing time implies that more experienced examiners may take longer to process applications, possibly due to handling complex cases or being thorough. This observation underscores the potential need for targeted training programs to enhance efficiency without compromising examination quality.

4. **Continuous Improvement:** The low R-squared values hint at other unexplored factors influencing processing times. The USPTO could benefit from ongoing efforts to identify these factors, using data analytics and refining predictive models to better understand and improve processing times.

5. **Policy and Strategic Planning:** These findings can guide policy and strategic decisions, especially concerning resource allocation, workload management, and operational efficiency. The non-significant impact of gender on processing times should reinforce the USPTO's commitment to gender equality and operational excellence.

In conclusion, the analysis indicates that examiner gender does not significantly impact the processing times of patent applications at the USPTO. This insight supports the continuation of gender equality initiatives and suggests a strategic focus on other factors that may enhance the efficiency and effectiveness of the examination process.