Exercise 3

Konstantin Volodin

2023-03-27

Summary Statistics

Workgroups 216 and 179 are evaluated. Workgroup 216 has 302 employees, and 179 has 829. Workgroup 216 is more male dominated with 73.5% vs 62.8% for 179. Both workgroup are mostly white and asian employees. With 179 having almost 80% white employees vs 60% for 216. Workgroup 216 generally has employees who worked less than 179

Table 1: Gender Distribution

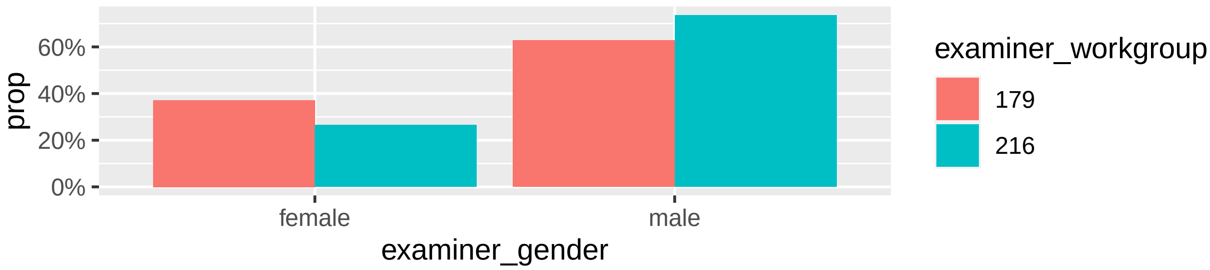
examiner_workgroup	female	male
179	37.15	62.85
216	26.49	73.51

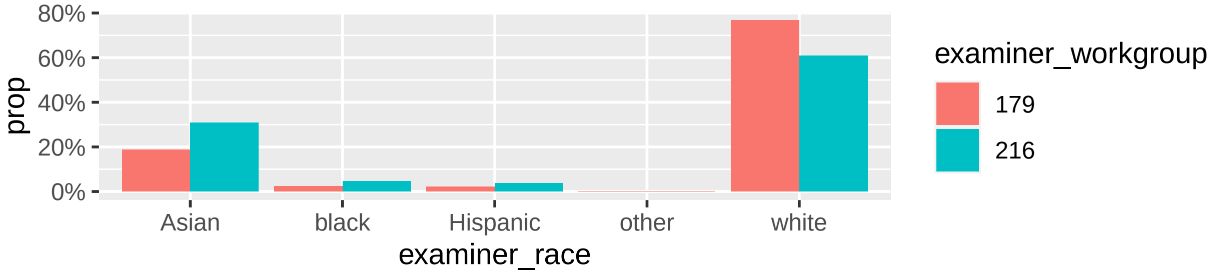
Table 2: Race Distribution

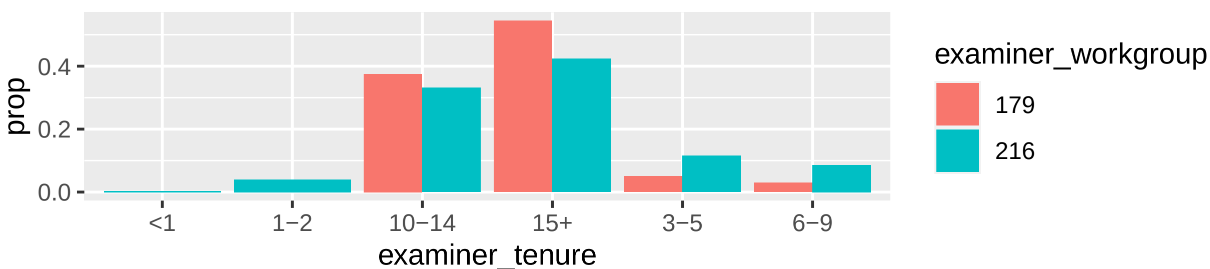
examiner_workgroup	Asian	black	Hispanic	other	white
179	18.70	2.29	2.17	0.12	76.72
216	30.79	4.64	3.64	NA	60.93

Table 3: Tenure Distribution

examiner_workgroup	10-14	15+	3-5	6-9	<1	1-2
179	37.52	54.40	5.07	3.02	NA	NA
216	33.11	42.38	11.59	8.61	0.33	3.97

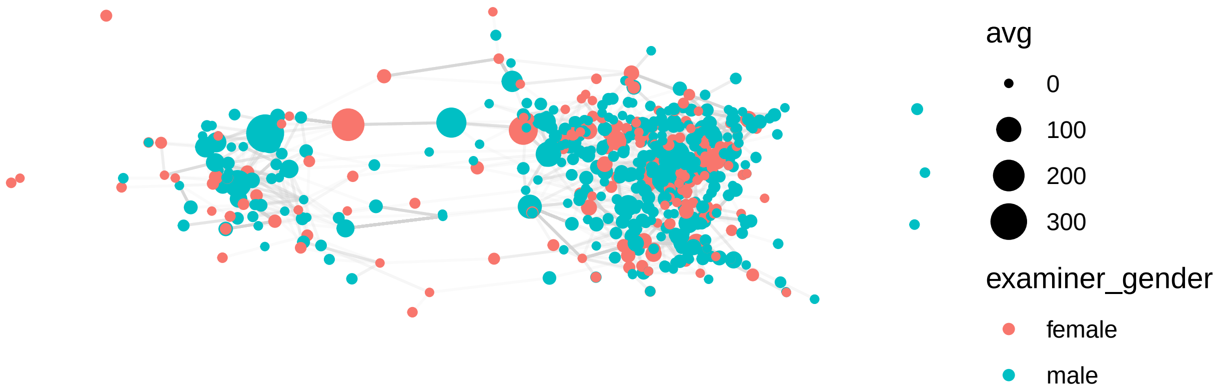


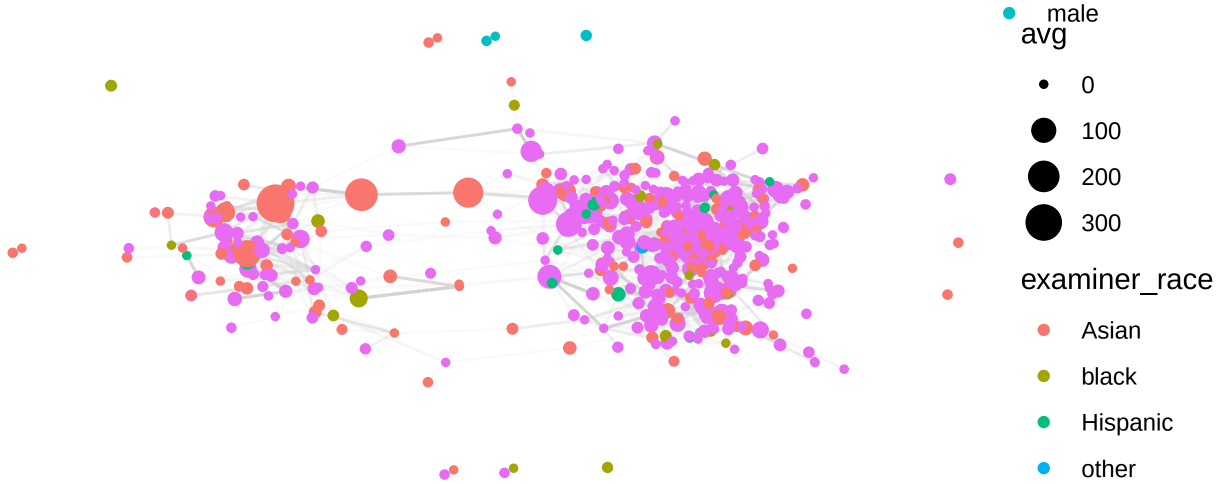




Network Visualization

There appears to be two distinct clusters. These clusters are likely base on workgroup however. Within each of the 2 clusters however there doesn't appear any segregation by gender or by race. This could be due to the non-dominant groups being too small to form their own cluster. Or it could be because the employees are interested in maintaining diverse groups.





Discussion

Table 4: Gender Centrality Scores

examiner_gender	top10_degree	top10_bet	mean_degree	mean_bet
female	30.78947	74.14260	5.010417	7.985249
male	32.53191	82.60917	5.674641	8.894336

Table 5: Race Centrality Scores

examiner_race	top10_degree	top10_bet	mean_degree	mean_bet
Asian	24.46154	161.73846	4.352941	15.92577
black	60.00000	5.00000	7.222222	0.50000
Hispanic	45.00000	0.00000	7.000000	0.00000
white	35.15556	67.25883	5.686230	6.94378

Code

```
### LOAD DATA
applications <- read_parquet(here('assignments','assignment_3',"app_data_clean.parquet"))
edges <- read_csv(here('assignments','assignment_3',"edges_sample.csv"))

### CLEAN DATA
```

```r
applications <- applications %>%
  select(-c('gender.y', 'race.y')) %>%
  rename(gender = gender.x, race = race.x) %>%
  mutate(tenure_years = tenure_days / 365) %>%
  mutate(tenure = case_when(
    tenure_years <= 1 ~ '<1',
    tenure_years <= 2 ~ '1-2',
    tenure_years <= 5 ~ '3-5',
    tenure_years <= 9 ~ '6-9',
    tenure_years <= 14 ~ '10-14',
    tenure_years <= 100 ~ '15+',
    TRUE ~ NA_character_
  ))


### WORKGROUPS
applications <- applications %>%
  mutate(examiner_workgroup = str_sub(examiner_art_unit, 1, -2))

### DROP NAs
applications <- applications %>% drop_na(gender, tenure, race)


### EXAMINER DATA
examiner_data <- applications %>%
  distinct(examiner_id, examiner_gender = gender,
           examiner_race = race, examiner_tenure = tenure)


### WORKGROUPS
examiner_subset <- applications %>%
  filter(examiner_workgroup %in% c(216, 179)) %>%
  distinct(examiner_id, examiner_workgroup) %>%
  left_join(examiner_data, by='examiner_id')


### COMPARE WORKGROUPS (STATISTICS)
t_gend <- examiner_subset %>% count(examiner_workgroup, examiner_gender) %>%
  group_by(examiner_workgroup) %>% mutate(freq = n / sum(n) * 100) %>%
  select(examiner_workgroup, examiner_gender, freq) %>%
  mutate(freq = round(freq, 2)) %>%
  pivot_wider(names_from = examiner_gender, values_from = freq)
t_race <- examiner_subset %>% count(examiner_workgroup, examiner_race) %>%
  group_by(examiner_workgroup) %>% mutate(freq = n / sum(n) * 100) %>%
  select(examiner_workgroup, examiner_race, freq) %>%
  mutate(freq = round(freq, 2)) %>%
  pivot_wider(names_from = examiner_race, values_from = freq)
t_tenure <- examiner_subset %>% count(examiner_workgroup, examiner_tenure) %>%
  group_by(examiner_workgroup) %>% mutate(freq = n / sum(n) * 100) %>%
  mutate(freq = round(freq, 2)) %>%
  select(examiner_workgroup, examiner_tenure, freq) %>%
  pivot_wider(names_from = examiner_tenure, values_from = freq)
```

```r
### COMPARE WORKGROUPS (PLOTS)
p_gend <- ggplot(examiner_subset, aes(x=examiner_gender, y=..prop..,
                                      fill=examiner_workgroup,
                                      group=examiner_workgroup)) +
  geom_bar(aes(), stat='count', position='dodge') +
  scale_y_continuous(labels = scales::percent_format())
p_race <- ggplot(examiner_subset, aes(x=examiner_race, y=..prop..,
                                      fill=examiner_workgroup,
                                      group=examiner_workgroup)) +
  geom_bar(aes(), stat='count', position='dodge') +
  scale_y_continuous(labels = scales::percent_format())
p_tenure <- ggplot(examiner_subset, aes(x=examiner_tenure, y=..prop..,
                                        fill=examiner_workgroup,
                                        group=examiner_workgroup)) +
  geom_bar(aes(), stat='count', position='dodge')


### CREATE NETWORK
edge_subset <- edges %>%
  filter(ego_examiner_id %in% examiner_subset$examiner_id &
           alter_examiner_id %in% examiner_subset$examiner_id) %>%
  drop_na() %>%
  select(to = ego_examiner_id, from = alter_examiner_id)
node_subset <- edge_subset %>%
  pivot_longer(cols=c('from','to')) %>%
  distinct(examiner_id = value) %>%
  left_join(examiner_data, on='examiner_id') %>%
  distinct(examiner_id, examiner_gender, examiner_race, examiner_tenure) %>%
  rename(name = examiner_id) %>%
  mutate(name = as.character(name))
network <- graph_from_data_frame(edge_subset, directed = TRUE) %>%
  as_tbl_graph() %>%
  left_join(node_subset, by='name')


### ESTIMATE METRICS
network <- network %>%
  mutate(degree = centrality_degree(),
         betweenness = centrality_betweenness()) %>%
  mutate(avg = (degree + betweenness)/2) %>%
  mutate(label = paste0(name, '\n',
                        'Degree: ',round(degree,2), '\n',
                        'Betweenness: ',round(betweenness,2), '\n',
                        'Avg: ',round(avg,2)))


### PLOT NETWORK
set.seed(1)
net_gender <- network %>%
  ggraph(layout="mds") +
  geom_edge_link(edge_colour = "#d3d3d3", alpha=0.1) +
  geom_node_point(aes(color=examiner_gender, size=avg)) +
  theme_void()
```

```r
set.seed(1)
net_race <- network %>%
  ggraph(layout="mds") +
  geom_edge_link(edge_colour = "#d3d3d3", alpha=0.1) +
  geom_node_point(aes(color=examiner_race, size=avg)) +
  theme_void()

### DISCUSSION
network_data <- network %>% as.data.frame() %>% as.tibble()

disc_gend_mean <- network_data %>%
  group_by(examiner_gender) %>%
  summarize(mean_degree = mean(degree),
            mean_bet = mean(betweenness))
disc_gend_top_degree <- network_data %>%
  arrange(desc(degree)) %>%
  group_by(examiner_gender) %>%
  top_frac(0.1, degree) %>%
  summarize(top10_degree = mean(degree))
disc_gend_top_bet <- network_data %>%
  arrange(desc(betweenness)) %>%
  group_by(examiner_gender) %>%
  top_frac(0.1, betweenness) %>%
  summarize(top10_bet = mean(betweenness))
disc_gend_top <- disc_gend_top_degree %>%
  left_join(disc_gend_top_bet, on='examiner_gender')
disc_gend <- disc_gend_top %>%
  left_join(disc_gend_mean, on='examiner_gender')



disc_race_mean <- network_data %>%
  group_by(examiner_race) %>%
  summarize(mean_degree = mean(degree),
            mean_bet = mean(betweenness))
disc_race_top_degree <- network_data %>%
  arrange(desc(degree)) %>%
  group_by(examiner_race) %>%
  top_frac(0.1, degree) %>%
  summarize(top10_degree = mean(degree))
disc_race_top_bet <- network_data %>%
  arrange(desc(betweenness)) %>%
  group_by(examiner_race) %>%
  top_frac(0.1, betweenness) %>%
  summarize(top10_bet = mean(betweenness))

disc_race_top <- disc_race_top_degree %>%
  left_join(disc_race_top_bet, on='examiner_race')
disc_race <- disc_race_top %>%
  left_join(disc_race_mean, on='examiner_race')
```