Darin Zlatarev #261081234

Professor Roman Galperin

ORGB 672 – Exercise 3

28 March 2023

Part 1: Loading and Preparation

Our goal is to analyze the various types of networks found in a USPTO dataset containing more than two million patent application entries (see Fig. 1).

```
> applications # Print the application table and the edge table
# A tibble: 2,018,477 × 21
   applica…¹ filing_d…² exami…³ exami…⁴ exami…⁵ exami…⁶ exami…⁷ uspc_…⁸ uspc_…⁹ paten…ˣ patent_i…ˣ abandon_…ˣ dispo…ˣ appl_…ˣ
   <chr>     <date>     <chr>   <chr>   <chr>     <dbl>   <dbl> <chr>   <chr>   <chr>   <date>     <date>     <chr>   <dbl>
 1 08284457  2000-01-26 HOWARD  JACQUE… V         96082    1764 508     273000  6521570 2003-02-18 NA         ISS       150
 2 08413193  2000-10-11 YILDIR… BEKIR   L         87678    1764 208     179000  6440298 2002-08-27 NA         ISS       250
 3 08531853  2000-05-17 HAMILT… CYNTHIA NA        63213    1752 430     271100  5607816 1997-03-04 NA         ISS       250
 4 08637752  2001-07-20 MOSHER  MARY    NA        73788    1648 530     388300  6927281 2005-08-09 NA         ISS       250
 5 08682726  2000-04-10 BARR    MICHAEL E         77294    1762 427     430100  NA      NA         2000-12-27 ABN       161
 6 08687412  2000-04-28 GRAY    LINDA   LAMEY     68606    1734 156     204000  6267836 2001-07-31 NA         ISS       150
 7 08716371  2004-01-26 MCMILL… KARA    RENITA    89557    1627 424     401000  NA      NA         NA         PEND      135
 8 08765941  2000-06-23 FORD    VANESSA L         97543    1645 424     001210  NA      NA         2001-08-22 ABN       161
 9 08776818  2000-02-04 STRZEL… TERESA  E         98714    1637 435     006000  NA      NA         2002-07-15 ABN       161
10 08809677  2002-02-20 KIM     SUN     U         65530    1723 210     645000  6858146 2005-02-22 NA         ISS       250
# … with 2,018,467 more rows, 7 more variables: appl_status_date <chr>, tc <dbl>, gender <chr>, race <chr>,
#   earliest_date <date>, latest_date <date>, tenure_days <dbl>, and abbreviated variable names ¹application_number,
#   ²filing_date, ³examiner_name_last, ⁴examiner_name_first, ⁵examiner_name_middle, ⁶examiner_id, ⁷examiner_art_unit,
#   ⁸uspc_class, ⁹uspc_subclass, ˣpatent_number, ˣpatent_issue_date, ˣabandon_date, ˣdisposal_type, ˣappl_status_code
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

Fig. 1

The table contains 21 columns, however it is missing a few key values such as the gender of the patent examiner and their race. Thus, we can write some code to estimate the gender based on the first name and their race based on their last name (see Fig. 2A, Fig. 2B). This code only assigns these values based on probabilities, so a few errors are imminent.

```
> examiner_names_gender # Estimate the gender based on the gender name
# A tibble: 1,822 × 3
   examiner_name_first gender proportion_female
   <chr>               <chr>              <dbl>
 1 AARON               male              0.0082
 2 ABDEL               male              0
 3 ABDOU               male              0
 4 ABDUL               male              0
 5 ABDULHAKIM          male              0
 6 ABDULLAH            male              0
 7 ABDULLAHI           male              0
 8 ABIGAIL             female            0.998
 9 ABIMBOLA            female            0.944
10 ABRAHAM             male              0.0031
# … with 1,812 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Fig. 2A

```
> examiner_race # Estimate the race based on the last name
# A tibble: 3,806 × 6
   surname    pred.whi pred.bla pred.his pred.asi pred.oth
   <chr>         <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
 1 HOWARD        0.597   0.295    0.0275   0.00690  0.0741
 2 YILDIRIM      0.807   0.0273   0.0694   0.0165   0.0798
 3 HAMILTON      0.656   0.239    0.0286   0.00750  0.0692
 4 MOSHER        0.915   0.00425  0.0291   0.00917  0.0427
 5 BARR          0.784   0.120    0.0268   0.00830  0.0615
 6 GRAY          0.640   0.252    0.0281   0.00748  0.0724
 7 MCMILLIAN     0.322   0.554    0.0212   0.00340  0.0995
 8 FORD          0.576   0.320    0.0275   0.00621  0.0697
 9 STRZELECKA    0.472   0.171    0.220    0.0825   0.0543
10 KIM           0.0169  0.00282  0.00546  0.943    0.0319
# … with 3,796 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Fig. 2B

Lastly, we need to calculate tenure (i.e. the length of time that an employee has been with the USPTO). We do that by finding the latest and earliest date of a patent application that they have handled then subtracting the difference to see how much time has elapsed (see Fig. 3).

```
# A tibble: 5,625 × 4
   examiner_id earliest_date latest_date tenure_days
         <dbl> <date>        <date>            <dbl>
 1       59012 2004-07-28    2015-07-24         4013
 2       59025 2009-10-26    2017-05-18         2761
 3       59030 2005-12-12    2017-05-22         4179
 4       59040 2007-09-11    2017-05-23         3542
 5       59052 2001-08-21    2007-02-28         2017
 6       59054 2000-11-10    2016-12-23         5887
 7       59055 2004-11-02    2007-12-26         1149
 8       59056 2000-03-24    2017-05-22         6268
 9       59074 2000-01-31    2017-03-17         6255
10       59081 2011-04-21    2017-05-19         2220
# … with 5,615 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Fig. 3

Part 2: Choosing Three Subsets

Our next step is to choose three different subset based on the first three digits of the examiners' art units. I randomly decided to choose units #161, #179, and #242. We then had to compare each subset's demographics using graphs and data. The first thing that came to my mind was to compare tenure days; perhaps some units were reserved for senior officers only and thus had more experienced personnel than the rest. Hence, I plotted three boxplot to visualize the differences, only to see that there is not much difference, nor variation between and within the units (see Fig. 4). The mean was around 6000 days in all three groups with very little variance, only a relatively small number of less experienced outliers (see Fig. 5). This means that all three units were full of experienced officers.

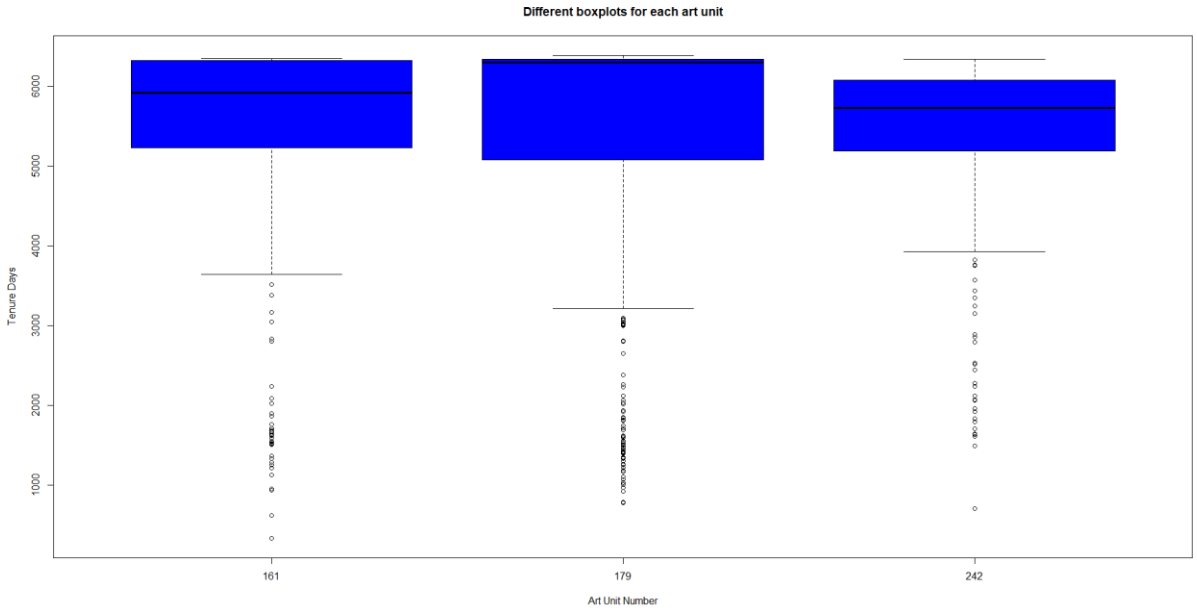

Fig. 4

```
> summary(APP161$tenure_days)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    330    5233    5918    5679    6327    6350    3731
> summary(APP179$tenure_days)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    774    5080    6304    5712    6342    6391    1058
> summary(APP242$tenure_days)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    703    5187    5726    5377    6082    6344     518
```

Fig. 5

Next, I tried to compared genders across all three subgroups, where there was a noticeable difference; it ranged from about 50% male in unit #161 to about 80% male in unit #242 (see Fig. 6 and Fig. 7).

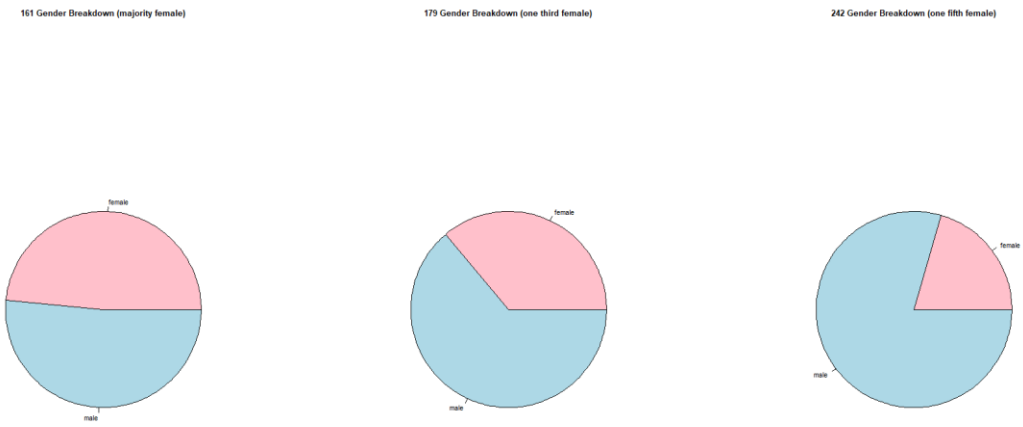

Fig. 6

```
> pie(table(APP161$gender), main="161 Gender Breakdown (majority female)", col=c("pink","light blue"))
> table(APP161$gender)

female   male
 37275  39554
> pie(table(APP179$gender), main="179 Gender Breakdown (one third female)", col=c("pink","light blue"))
> table(APP179$gender)

female   male
 43783  77344
> pie(table(APP242$gender), main="242 Gender Breakdown (one fifth female)", col=c("pink","light blue"))
> table(APP242$gender)

female   male
  4968  19187
> par(mfrow=c(1,4)) # Reset the size
> pie(table(APP161$race), main = "161 Race Breakdown")
> table(APP161$race)
```

Fig. 7

Lastly, I compared race across all three groups, as well as in all three groups together to see if there are any interesting patterns. Units #161 and #179 are predominantly white at about a 75% absolute majority (see Fig. 8). Unit #242 is the most diverse one by far; whites only make a 45% relative majority with Asians occupying about a third. The proportion of blacks and Hispanics is also significantly larger in unit #242. Oddly enough, unit #179 is the only one with people whose race is classified as “other” as there are 24 such individuals (see Fig. 9).

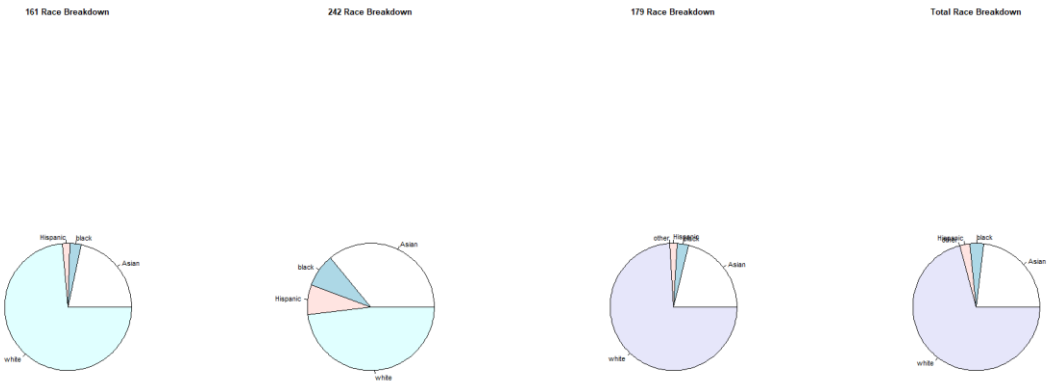

Fig. 8

```
> par(mfrow=c(1,4)) # Reset the size
> pie(table(APP161$race), main = "161 Race Breakdown")
> table(APP161$race)

   Asian    black Hispanic    white
   19528     2452     1843    65972
> pie(table(APP242$race), main = "242 Race Breakdown")
> table(APP242$race)

   Asian    black Hispanic    white
   10874     2530     2319    14520
> pie(table(APP179$race), main = "179 Race Breakdown")
> table(APP179$race)

   Asian    black Hispanic    other    white
   28335     3771     2449       24    98845
> pie(table(APPBIG$race), main = "Total Race Breakdown") # Include a pie chart for all subsets together
> table(APPBIG$race)

   Asian    black Hispanic    other    white
   58737     8753     6611       24   179337
> applications # Print the application table and the edge table
```

Fig. 9

Part 3: Plotting the Networks

The last step is to plot the actual network. I plotted two plots: one based on gender and the other one based on race. In both cases I colored the nodes accordingly and made node size be determined by tenure days (see Fig. 10 and Fig. 11). The reason why I did this was because I thought that more experienced individuals would have more betweenness and closeness centrality, however that was not always the case since some individuals with less experience got to act as a bridge between subgroups.

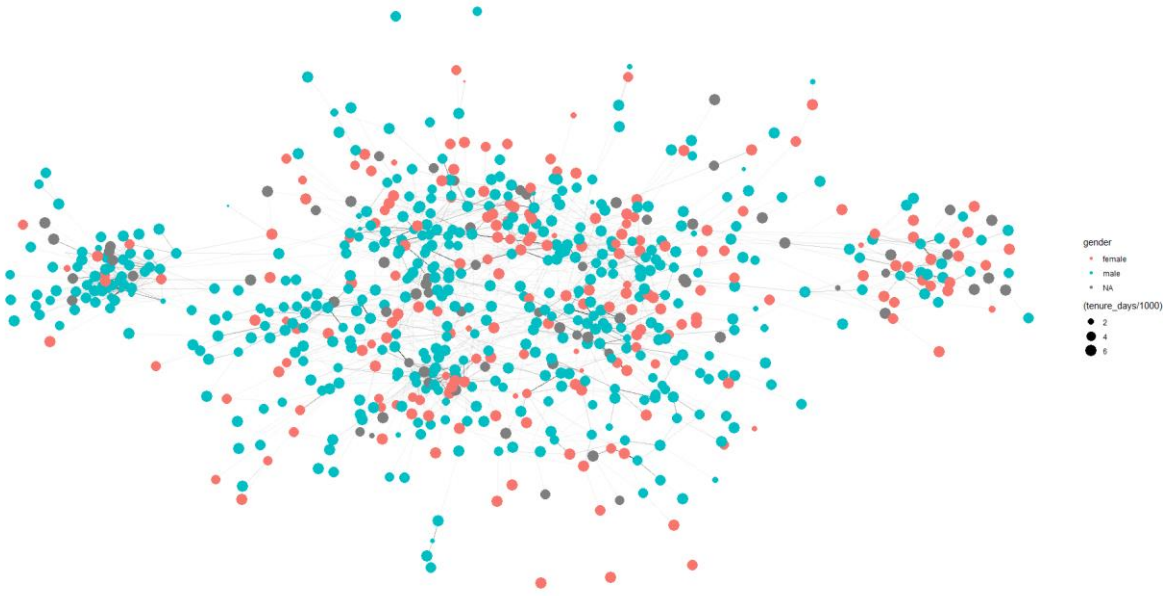

Fig. 10

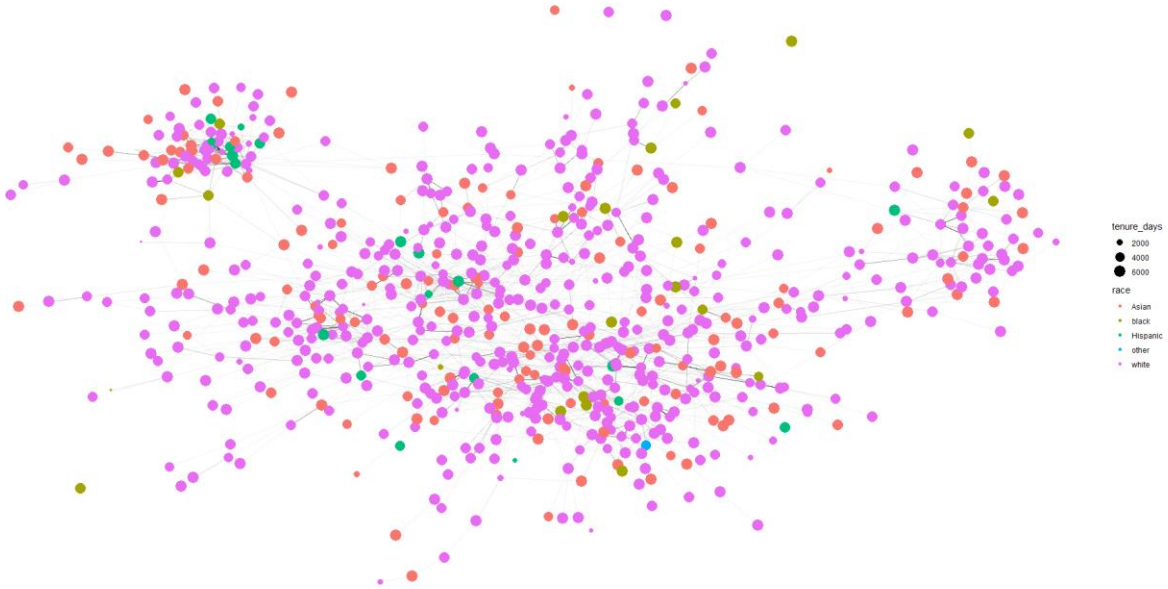
tenure_days
2000
4000
6000
race
Asian
black
Hispanic
other
white

Fig. 11