STAT 231: Problem Set 7B

Majd Rouhana

due by 5 PM on Friday, October 30

This homework assignment is designed to help you futher ingest, practice, and expand upon the material covered in class over the past week(s). You are encouraged to work with other students, but all code and text must be written by you, and you must indicate below who you discussed the assignment with (if anyone).

Steps to proceed:

1. In RStudio, go to File > Open Project, navigate to the folder with the course-content repo, select the course-content project (course-content.Rproj), and click "Open"
2. Pull the course-content repo (e.g. using the blue-ish down arrow in the Git tab in upper right window)
3. Copy ps7B.Rmd from the course repo to your repo (see page 6 of the GitHub Classroom Guide for Stat231 if needed)
4. Close the course-content repo project in RStudio
5. Open YOUR repo project in RStudio
6. In the ps7B.Rmd file in YOUR repo, replace "YOUR NAME HERE" with your name
7. Add in your responses, committing and pushing to YOUR repo in appropriate places along the way
8. Run "Knit PDF"
9. Upload the pdf to Gradescope. Don't forget to select which of your pages are associated with each problem. *You will not get credit for work on unassigned pages (e.g., if you only selected the first page but your solution spans two pages, you would lose points for any part on the second page that the grader can't see).*

If you discussed this assignment with any of your peers, please list who here:

ANSWER:

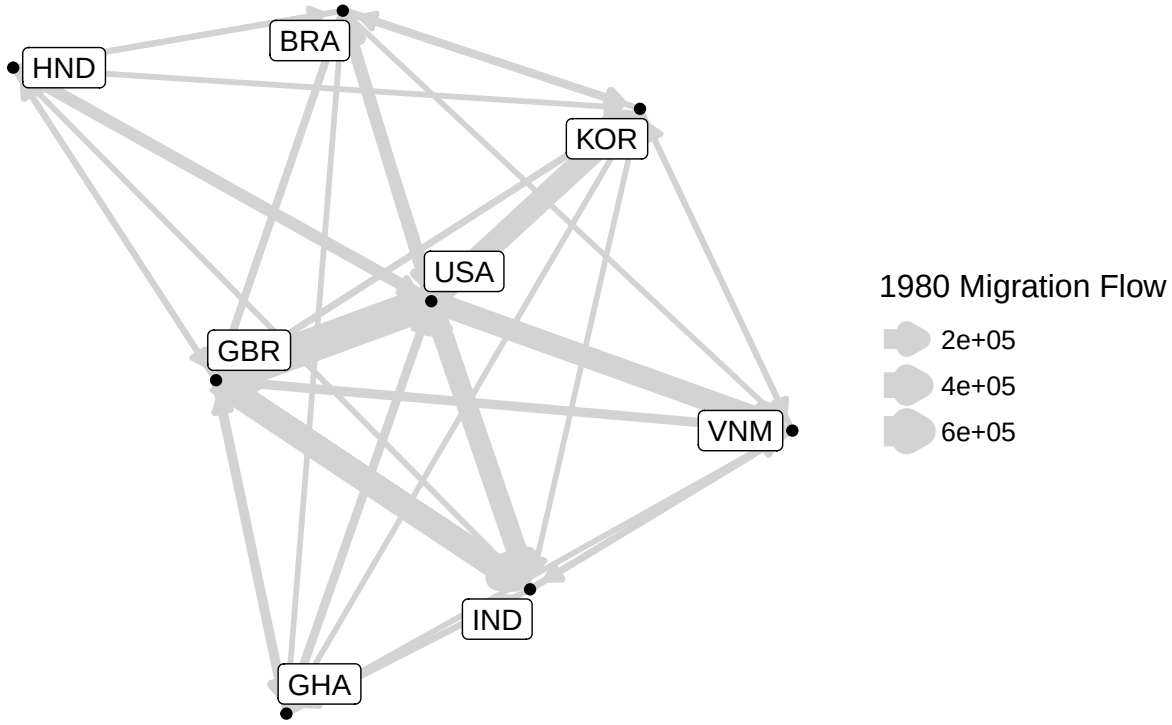
1. More Migration

1a. Consider migration between the following countries: Brazil, Ghana, Great Britain, Honduras, India, South Korea, United States, and Vietnam. Compare the TOTAL (males + females) migration between these countries over time. In separate (directed) graphs for 1980 and 2000, visualize the network for the these countries with edge width and/or edge color corresponding to migration flow size. Interpret the two graphs – what *information in context* do they convey?

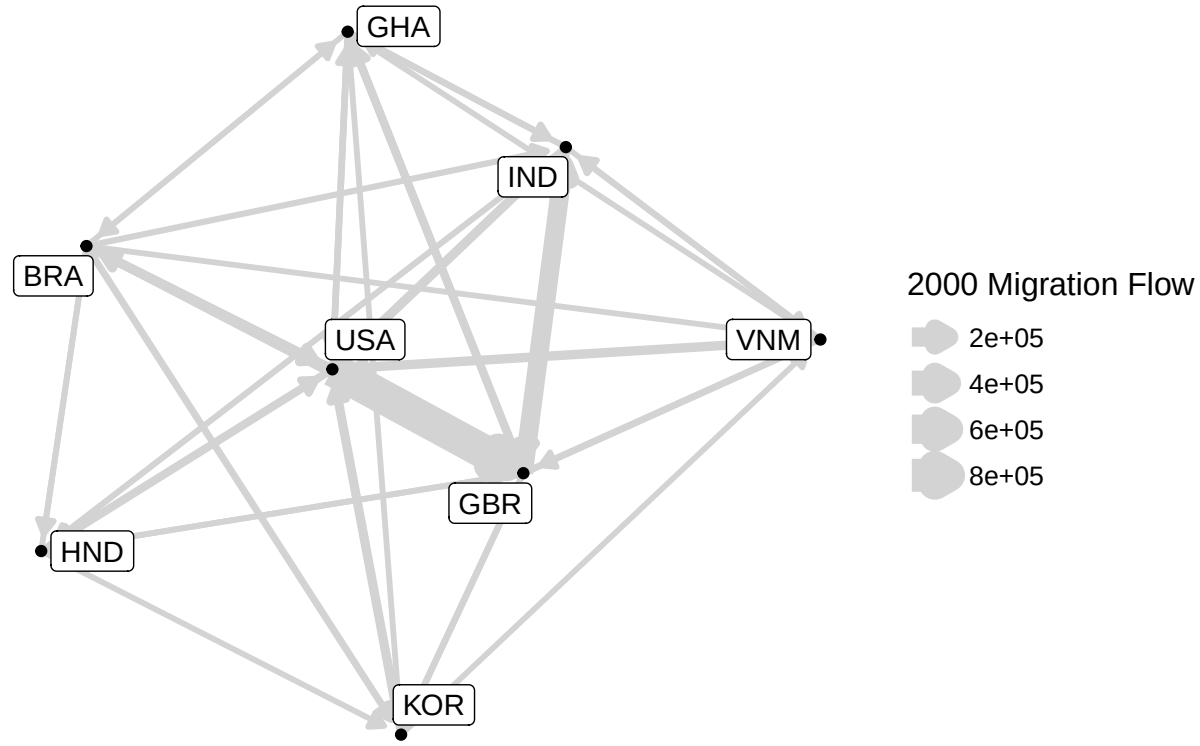
> ANSWER: From both graphs, US appear to be by far the most popular for migration destinations, with Great Britain in second place. Conversely, it seems that Vietnam is the least-popular migration destination, with the second least-popular being Honduras. Looking at these trends, the graphs seem to tell us that it was more common for people to migrate to first-world countries than third-world countries in the years 1980 and 2000.

```
path_in <- "/home/class22/mrouhana22/git/majd_STAT231/homework"
MigrationFlows <- read_csv(paste0(path_in,"/MigrationFlows.csv"))

countries <- c("BRA","GBR","GHA","HND","IND","KOR", "USA", "VNM")

# need migration overall:
# do some prelim data wrangling to combine numbers for males + females

MigrationFlows1980 <- MigrationFlows %>%
  select(-c(sex, Y2000, Y1990, Y1970, Y1960)) %>%
  group_by(destcode, origincode) %>%
  summarize(Y1980 = sum(Y1980)) %>%
  filter(Y1980 > 0, destcode %in% countries & origincode %in% countries)

MigrationFlows2000 <- MigrationFlows %>%
  select(-c(sex, Y1990, Y1980, Y1970, Y1960)) %>%
  group_by(destcode, origincode) %>%
  summarize(Y2000 = sum(Y2000)) %>%
  filter(Y2000 > 0, destcode %in% countries & origincode %in% countries)

migration1980 <- graph_from_data_frame(MigrationFlows1980
                                       , directed = TRUE)

migration2000 <- graph_from_data_frame(MigrationFlows2000
                                       , directed = TRUE)

migration_network80 <- ggnetwork(migration1980)

migration_network00 <- ggnetwork(migration2000)

ggplot(data = migration_network80
       , aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_edges(arrow = arrow(type = "closed", length = unit(6, "pt"))
             , color = "lightgray"
             , aes(size = Y1980)) +
  geom_nodes() +
  geom_nodelabel_repel(aes(label = name)) +
  theme_blank() +
  ggtitle("Migration Flow for Select Countries in 1980") +
  labs(size = "1980 Migration Flow")
```

Migration Flow for Select Countries in 1980




```
ggplot(data = migration_network00
       , aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_edges(arrow = arrow(type = "closed", length = unit(6, "pt"))
             , color = "lightgray"
             , aes(size = Y2000)) +
  geom_nodes() +
  geom_nodelabel_repel(aes(label = name)) +
  theme_blank() +
  ggtitle("Migration Flow for Select Countries in 2000") +
  labs(size = "2000 Migration Flow")
```



1b. Compute the *unweighted* in-degree for Brazil in this network from 2000, and the *weighted* in-degree for Brazil in this network from 2000. In 1-2 sentences, interpret these numbers in context (i.e., without using the terms "in-degree" or "weighted").

> ANSWER: The unweighted in-degree for Brazil is 4, while the weighted in-degree is 18050. In context, this means that there is a relatively low number of different countries that people come from that are migrating into Brazil compared to the other selected countries.

```
igraph::degree(migration2000, mode = "in")
```

```
## BRA GBR GHA HND IND KOR USA VNM
##   4   7   4   4   6   5   7   6
```

```
strength(migration2000, weights = E(migration2000)$Y2000
         , mode = "in")
```

```
##    BRA    GBR    GHA    HND    IND    KOR    USA    VNM
##  18050 899064   6926   7151 206251  15501 145567  16230
```

1c. Among these same countries, identify the top 5 countries *of origin* and *of destination* (separately) in 1980 using (weighted) degree centrality. Interpret this information.

> ANSWER: The top 5 countries of origin are the United States, Great Britain, Brazil, India and Korea. The top 5 countries of destination are Great Britain, India, Korea, Vietnam and the United States. From this information, it seems that countries with higher populations seem to make both lists, as more people are coming in and out of them. Great Britain also seems to be the most popular destination in 1980 proportionally, even though the United States has the most number of people migrating into it in 1980. The United States also seems to have, proportionally, the most people migrating out of the country in 1980. Also, it seems that countries with smaller populations have a higher percentage of people migrating into the country than out. For example, while Vietnam had the least number of people who migrated into the country, its

weighted in-degree of 278247 is much higher than its weighted out-degree of 743. This indicates that much more people, proportionally are migrating into Vietnam than those who are migrating out.

```
strength(migration1980, weights = E(migration1980)$Y1980)
```

```
##     BRA     GBR     GHA     HND     IND     KOR     USA     VNM
##   79544 1370224   29854   44692  646972  326491 1848395  278990
```

```
strength(migration1980, weights = E(migration1980)$Y1980
         , mode = "in")
```

```
##    BRA    GBR    GHA    HND    IND    KOR    USA    VNM
##  53035 812225  27505  43500 631220 321966 144883 278247
```

```
strength(migration1980, weights = E(migration1980)$Y1980
         , mode = "out")
```

```
##     BRA     GBR     GHA     HND     IND     KOR     USA     VNM
##   26509  557999    2349    1192   15752    4525 1703512     743
```

1d. Among these same countries, identify the top 5 countries *of origin* and *of destination* (separately) in 2000 using (weighted) degree centrality. Interpret this information.

ANSWER: The top 5 countries of origin are the United States, Great Britain, Brazil, India and Ghana. The top 5 countries of destination are Great Britain, India, the United States, Brazil and Vietnam. The results are similar to the last question, as Great Britain seems to be the most popular destination in 2000 proportionally, while the United States seems to proportionately have the most people migrating out of it in 2000. Something different from this year is that more people, proportionately, seem to be migrating out of Ghana compared to people migrating in. In 1980, it was very lopsided, with much much more people migrating in than out. This is interesting to see. Another thing to note is that India and South Korea seems to both be much less-popular destinations in 2000 than 1980.

```
strength(migration2000, weights = E(migration2000)$Y2000)
```

```
##     BRA     GBR     GHA     HND     IND     KOR     USA     VNM
##   38935 1220029   15513    9004  226493   22374 1080364   16768
```

```
strength(migration2000, weights = E(migration2000)$Y2000
         , mode = "in")
```

```
##    BRA    GBR    GHA    HND    IND    KOR    USA    VNM
##  18050 899064   6926   7151 206251  15501 145567  16230
```

```
strength(migration2000, weights = E(migration2000)$Y2000
         , mode = "out")
```

```
##    BRA    GBR    GHA    HND    IND    KOR    USA    VNM
##  20885 320965   8587   1853  20242   6873 934797    538
```

1e. What is the diameter of this network in 2000? In 1-2 sentences, interpret this value.

ANSWER: The diameter of this network in 2000 is 2. This means that the shortest path between any 2 countries is 2 countries away.

```
diameter(migration2000, directed = FALSE)
```

```
## [1] 2
```

1f. What is the density of this network in 2000? In 1-2 sentences, interpret this value.

ANSWER: The density of this network in 2000 is about 76.79%. This indicates that the network is very connected.

```
graph.density(migration2000)
```

```
## [1] 0.7678571
```

2. Love Actually (OPTIONAL PRACTICE)

This problem is *optional* and will not be graded, but is given to provide additional practice interpreting networks and as another real-world example of network analysis that might be intriguing to film buffs.

Consider the figure "The Two Londons of 'Love Actually' " in this FiveThirtyEight article.

2a. Based on this figure, is the network connected? In 1-2 sentences, please explain.

ANSWER:

2b. Based on the figure, what is the (unweighted) degree for Emma Thompson? What is the (unweighted) degree for Keira Knightley? Explain what these values mean for these characters.

ANSWER:

2c. Based on the figure, for whom would the (unweighted) betweenness centrality measure be higher: Colin Firth or Hugh Grant? Explain what this implies.

ANSWER: