STAT 231: Problem Set 7B

Sean Wei

due by 5 PM on Friday, October 30

This homework assignment is designed to help you further 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:

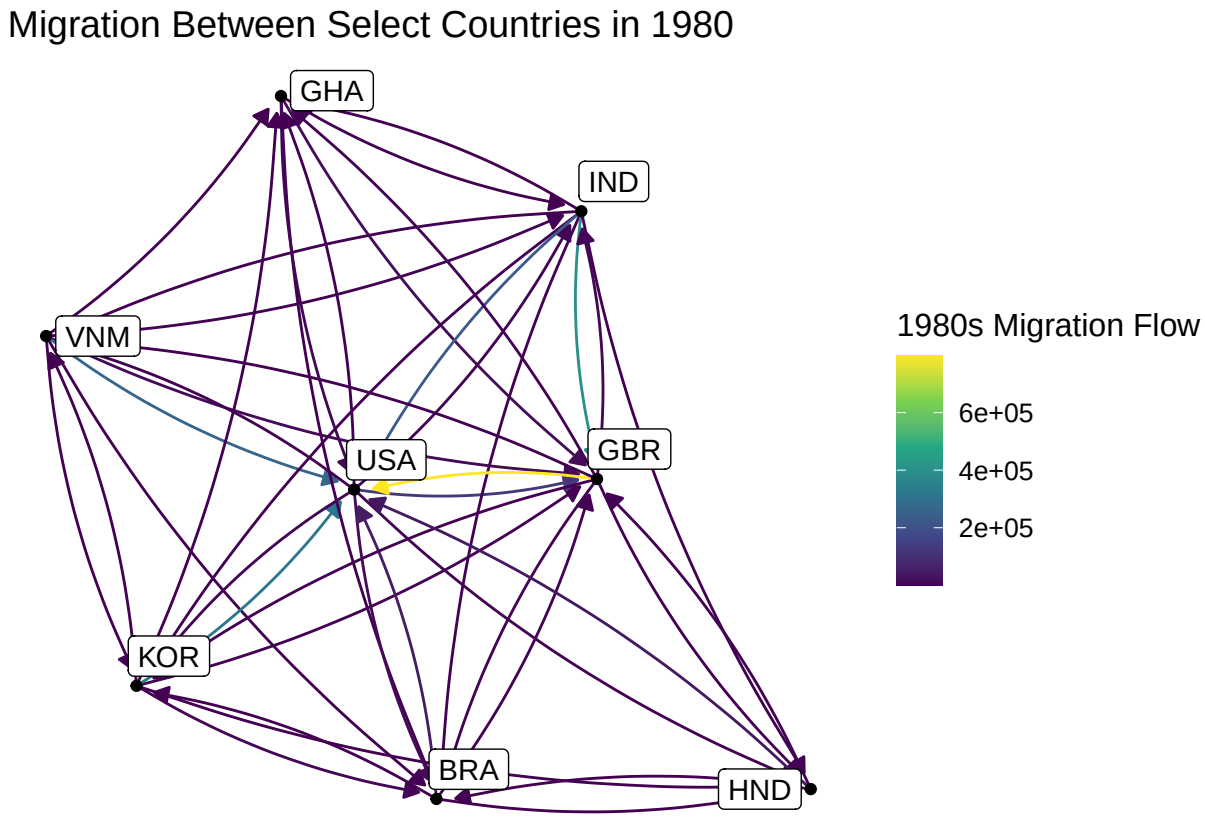
ANSWER:

1. More Migration

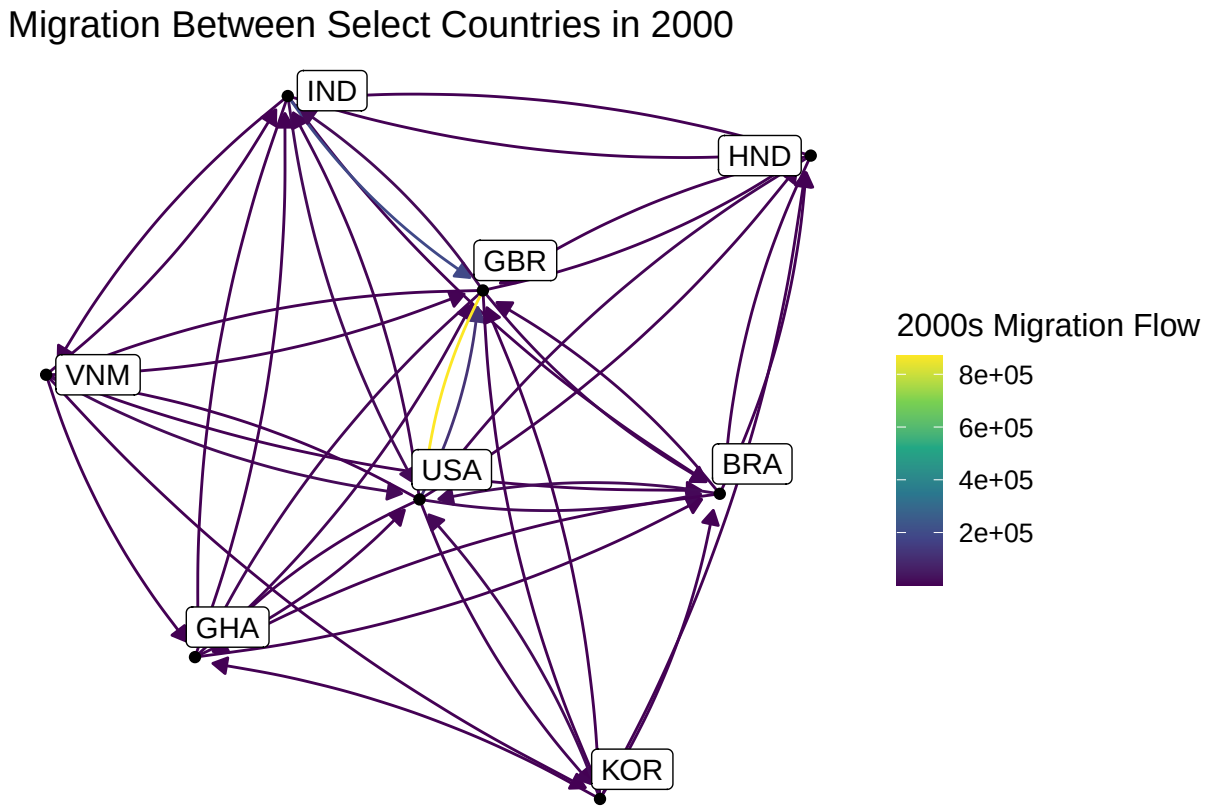
1a. Consider migration between the following countries: Brazil, Ghana, Great Britain, Honduras, India, South Korea, United States, and Vietnam. Compare the TOTAL (males + females) migration between these countries over time. In separate (directed) graphs for 1980 and 2000, visualize the network for these 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: These two graphs convey information about the total number of people who have migrated between BRA, GBR, GHA, HND, IND, KOR, USA, and VNM in 1980 and 2000. From the graphs, it is clear that more migration occurred between these countries in 1980 than 2000. Also, the most migrations in both years were from GBR to USA.

```
path_in <- "/Users/seanwei/Desktop/STAT231-swei1999/data"
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

# 1980s Migration Flow
MigrationFlows1980 <- MigrationFlows %>%
  filter(Y1980 > 0) %>%
  select(origincode, destcode, Y1980) %>%
  filter(destcode %in% countries & origincode %in% countries) %>%
  group_by(origincode, destcode) %>%
  summarise(Y1980 = sum(Y1980))

MigrationFlows1980 <- graph_from_data_frame(MigrationFlows1980, directed = TRUE)

migration_network_1980 <- ggnetwork(MigrationFlows1980)

ggplot(data = migration_network_1980, aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_edges(curvature = 0.1,
             arrow = arrow(type = "closed", length = unit(6, "pt")),
             aes(color = Y1980)) +
  geom_nodes() +
  geom_nodelabel_repel(aes(label = name)) +
  theme_blank() +
  ggtitle("Migration Between Select Countries in 1980") +
  labs(color = "1980s Migration Flow") +
  scale_color_continuous(type = "viridis")
```




```
# 2000s Migration Flow
MigrationFlows2000 <- MigrationFlows %>%
  filter(Y2000 > 0) %>%
  select(origincode, destcode, Y2000) %>%
  filter(destcode %in% countries & origincode %in% countries) %>%
  group_by(origincode, destcode) %>%
  summarise(Y2000 = sum(Y2000))

MigrationFlows2000 <- graph_from_data_frame(MigrationFlows2000, directed = TRUE)

migration_network_2000 <- ggnetwork(MigrationFlows2000)

ggplot(data = migration_network_2000, aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_edges(curvature = 0.1,
             arrow = arrow(type = "closed", length = unit(6, "pt")),
             aes(color = Y2000)) +
  geom_nodes() +
  geom_nodelabel_repel(aes(label = name)) +
  theme_blank() +
  ggtitle("Migration Between Select Countries in 2000") +
  labs(color = "2000s Migration Flow") +
  scale_color_continuous(type = "viridis")
```



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 in this network from 2000 is 7. This means that there were 7 other countries that people left from to migrate to Brazil. In addition, the weighted in-degree for Brazil in this network from 2000 is 20885, which means that 20885 people migrated to Brazil in 2000.

```
V(MigrationFlows2000)$degree <- igraph::degree(MigrationFlows2000, mode = "in")
V(MigrationFlows2000)$wtdegree <- strength(MigrationFlows2000,
                                    weights = E(MigrationFlows2000)$Y2000, mode = "in")
stats <- data.frame(name = V(MigrationFlows2000)$name,
                    degree = V(MigrationFlows2000)$degree,
                    wtdegree = V(MigrationFlows2000)$wtdegree)
stats %>%
  filter(name == "BRA")
```

```
##   name degree wtdegree
## 1  BRA      7    20885
```

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 in 1980 were GBR, IND, KOR, VNM, and USA. This means that these were the top 5 most popular countries to migrate from in 1980 of the selected

countries. The top 5 countries of destination in 1980 were USA, GBR, BRA, IND, and KOR. This means that these were the top 5 most popular countries to migrate to in 1980 of the selected countries.

```
# of origin
head(sort.int(strength(MigrationFlows1980, weights = E(MigrationFlows1980)$Y1980, mode = "out"),
              decreasing = TRUE), 5)
```

```
##    GBR    IND    KOR    VNM    USA
## 812225 631220 321966 278247 144883
```

```
# of destination
head(sort.int(strength(MigrationFlows1980, weights = E(MigrationFlows1980)$Y1980, mode = "in"),
              decreasing = TRUE), 5)
```

```
##     USA     GBR     BRA     IND     KOR
## 1703512  557999   26509   15752    4525
```

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 in 2000 were GBR, IND, USA, BRA, and VNM. This means that these were the top 5 most popular countries to migrate from in 2000 of the selected countries. The top 5 countries of destination in 2000 were USA, GBR, BRA, IND, and GHA. This means that these were the top 5 most popular countries to migrate to in 2000 of the selected countries.

```
# of origin
head(sort.int(strength(MigrationFlows2000, weights = E(MigrationFlows2000)$Y2000, mode = "out"),
              decreasing = TRUE), 5)
```

```
##    GBR    IND    USA    BRA    VNM
## 899064 206251 145567  18050  16230
```

```
# of destination
head(sort.int(strength(MigrationFlows2000, weights = E(MigrationFlows2000)$Y2000, mode = "in"),
              decreasing = TRUE), 5)
```

```
##    USA    GBR    BRA    IND    GHA
## 934797 320965  20885  20242   8587
```

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, meaning that the longest geodesic (shortest path) between any two countries for migration is 2 countries away.

```
diameter(MigrationFlows2000, 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 ~76.8%, which tells us that the network is largely concentrated. This means that most countries are connected to many others.

```
graph.density(MigrationFlows2000)
```

```
## [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: