STAT 231: Problem Set 7B

Jack Dove

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:

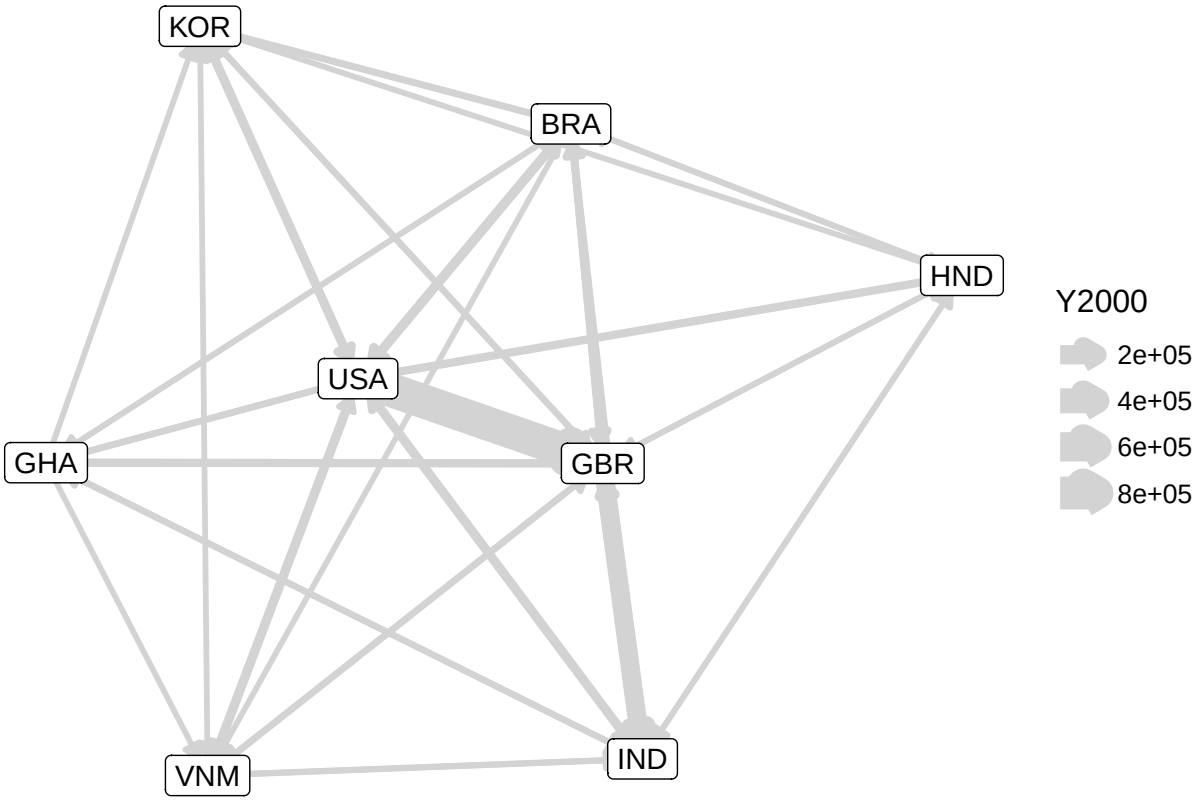
1. More Migration

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?

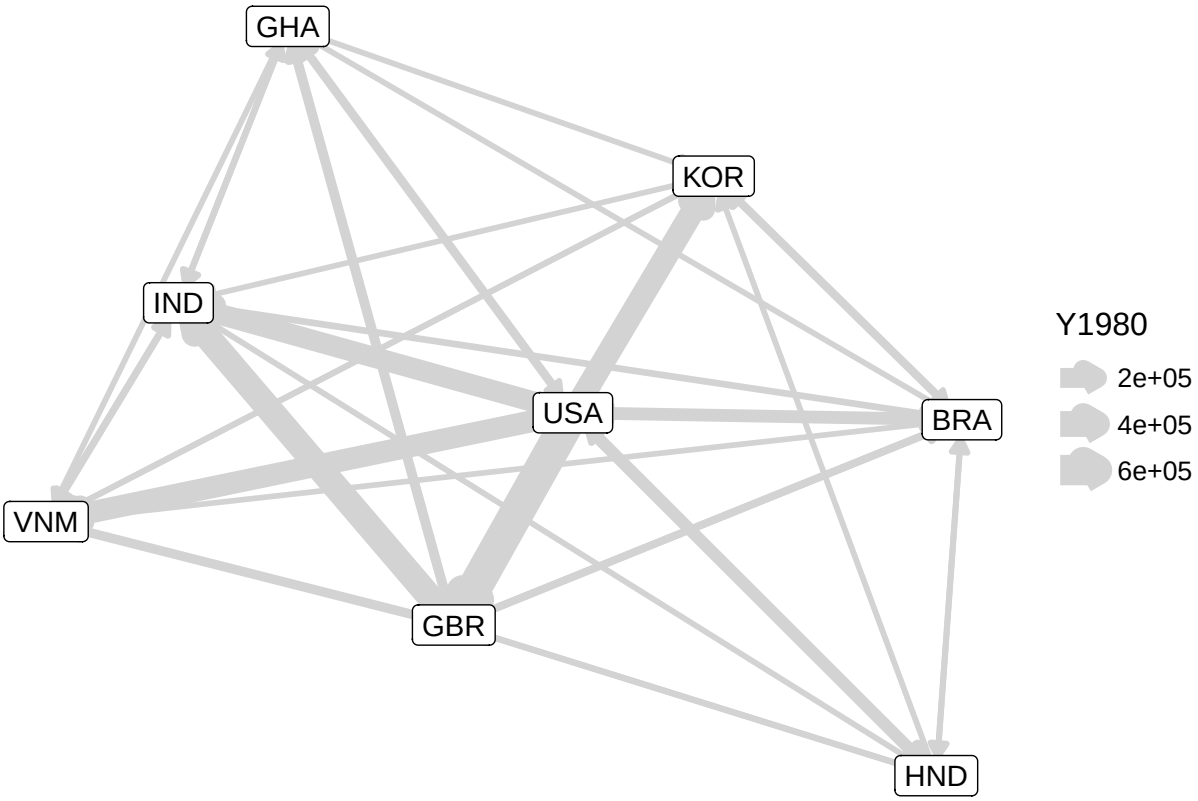
> ANSWER: While in 1980 the immigration trends were more evenly distributed, they are dominated by the US and the UK in 2000. The US and UK are central figures in the 2000 network, and Vietnam and India are more involved in the 1980 network.

```
library(csvread)

path_in <- "~/Desktop/Data Science/Stat231JackDove/Homeworks"
MigrationFlows <- read_csv(paste0(path_in,"/MigrationFlows.csv"))

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

migration <- MigrationFlows %>%
  filter(destcode %in% countries) %>%
  filter(origincode %in% countries) %>%
  unite(col = "origin_to_destination", origincode, destcode, sep="_") %>%
  pivot_wider(names_from = sex, values_from = c(Y2000, Y1990, Y1980, Y1970, Y1960)) %>%
  mutate(Y2000 = Y2000_Male + Y2000_Female) %>%
  mutate(Y1990 = Y1990_Male + Y1990_Female) %>%
  mutate(Y1980 = Y1980_Male + Y1980_Female) %>%
  mutate(Y1970 = Y1970_Male + Y1970_Female) %>%
  mutate(Y1960 = Y1960_Male + Y1960_Female) %>%
  select(origin_to_destination, Y2000, Y1990, Y1980, Y1970, Y1960) %>%
  separate(origin_to_destination, into = c("origincode", "destcode"))

migration2000 <- migration %>%
  select(destcode, origincode, Y2000) %>%
  filter(Y2000 > 0) %>%
  graph_from_data_frame(directed=TRUE)

migration2000network <- migration2000 %>%
  ggnetwork()

migration1980 <- migration %>%
  select(destcode, origincode, Y1980) %>%
  filter(Y1980 > 0) %>%
  graph_from_data_frame(directed=TRUE)

migration1980network <- migration1980 %>%
  ggnetwork()

#2000 Immigration Network
ggplot(data = migration2000network
       , 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(aes(label = name)) +
  theme_blank()
```




```
#1980 Immigration Network
ggplot(data = migration1980network
       , 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(aes(label = name)) +
  theme_blank()
```



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: While Brazil doesn't have that many connections to other countries, its individual connections are with countries with lots of connections, suggesting that Brazil does not have low but somewhat high immigration levels (people moving to Brazil). This trend is displayed by a low degree (T-6th/8) when considering only numbers of edges, but a medium-to-high degree (4th/8) when adjusting for the degrees of other connected edges.

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

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

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

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

```
#Brazil
#Unweighted: 4
#Weighted: 18050
```

1c. Among these same countries, idenitify 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 the US, the UK, Brazil, India, and Korea, while the top 5 countries of destination in 1980 were the UK, India, Korea, Vietnam, and the US. While the UK, India, Korea, and the US made both lists, only the US saw a higher out-degree, meaning it was more central from an emigration standpoint than from an immigration viewpoint.

```
# in
indegree1980 <- strength(migration1980, weights = E(migration1980)$Y1980
         , mode = "in")

# out
outdegree1980 <- strength(migration1980, weights = E(migration1980)$Y1980
         , mode = "out")

head(sort(indegree1980, decreasing = TRUE), n=5)
```

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

```
head(sort(outdegree1980, decreasing = TRUE), n=5)
```

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

1d. Among these same countries, idenitify 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 the US, the UK, Brazil, India, and Ghana, while the top 5 countries of destination in 2000 were the UK, India, US, Brazil, and the Vietnam. In contrast to 1980, Ghana overtook Korea in immigration centrality, meaning its immigrants came from countries who also had lots of immigrants. Meanwhile, India had similar levels of centrality in both immigration and emigration.

```
# in
indegree2000 <- strength(migration2000, weights = E(migration2000)$Y2000
         , mode = "in")

# out
outdegree2000 <- strength(migration2000, weights = E(migration2000)$Y2000
         , mode = "out")

head(sort(indegree2000, decreasing = TRUE), n=5)
```

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

```
head(sort(outdegree2000, decreasing = TRUE), n=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 is 2, meaning that the longest geodesic between two countries goes through one country (at the most).

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

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

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

> ANSWER: The density of the network is 0.76, suggesting that global immigration is very connected. If countries were isolationist, density would be zero, and if it were perfectly connected, density would be one, which means that the 0.76 value points to a well linked immigration network.

```
# density = #edges / #possibleedges

#number of edges
edges <- ecount(migration2000)

#number of possible edges (n *(n-1))
possibleedges <- (8)*(8-1)

edges/possibleedges
```

```
## [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: