# STAT 231: Problem Set 7B

Mike Santos

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: The data indicates that there was a significant decrease in migration between these countries from the 1980s to the 2000, the countries had especially been migrating to the USA during the 80's.

```
library(dplyr)
library(igraph)
library(network)
MigrationFlows <- read_csv("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
MigrationFlows= select(MigrationFlows, c(-sex))
Migration_Base = MigrationFlows %>% select(origincode, destcode, Y2000, Y1990, Y1980)
Migration = Migration_Base %>% filter(destcode %in% countries & origincode %in% countries)
migration = graph_from_data_frame(Migration, directed = TRUE)

# vertices
V(migration)
```

```
## + 8/8 vertices, named, from 76db35f:
## [1] BRA GHA HND IND KOR GBR USA VNM
```

```
V(migration)$Y2000
```

```
## NULL
```

```
V(migration)$Y1980
```

```
## NULL
```

```
vcount(migration)
```

```
## [1] 8
```

```
# edges
E(migration)
```

```
## + 128/128 edges from 76db35f (vertex names):
##  [1] BRA->GHA GHA->GHA HND->GHA IND->GHA KOR->GHA GBR->GHA USA->GHA VNM->GHA
##  [9] BRA->BRA GHA->BRA HND->BRA IND->BRA KOR->BRA GBR->BRA USA->BRA VNM->BRA
```

```
## [17] BRA->HND GHA->HND HND->HND IND->HND KOR->HND GBR->HND USA->HND VNM->HND
## [25] BRA->IND GHA->IND HND->IND IND->IND KOR->IND GBR->IND USA->IND VNM->IND
## [33] BRA->KOR GHA->KOR HND->KOR IND->KOR KOR->KOR GBR->KOR USA->KOR VNM->KOR
## [41] BRA->GBR GHA->GBR HND->GBR IND->GBR KOR->GBR GBR->GBR USA->GBR VNM->GBR
## [49] BRA->USA GHA->USA HND->USA IND->USA KOR->USA GBR->USA USA->USA VNM->USA
## [57] BRA->VNM GHA->VNM HND->VNM IND->VNM KOR->VNM GBR->VNM USA->VNM VNM->VNM
## [65] BRA->GHA GHA->GHA HND->GHA IND->GHA KOR->GHA GBR->GHA USA->GHA VNM->GHA
## [73] BRA->BRA GHA->BRA HND->BRA IND->BRA KOR->BRA GBR->BRA USA->BRA VNM->BRA
## + ... omitted several edges
```

```
E(migration)$destcode
```

```
## NULL
```

```
E(migration)$origincode
```

```
## NULL
```

```
ecount(migration)
```

```
## [1] 128
```

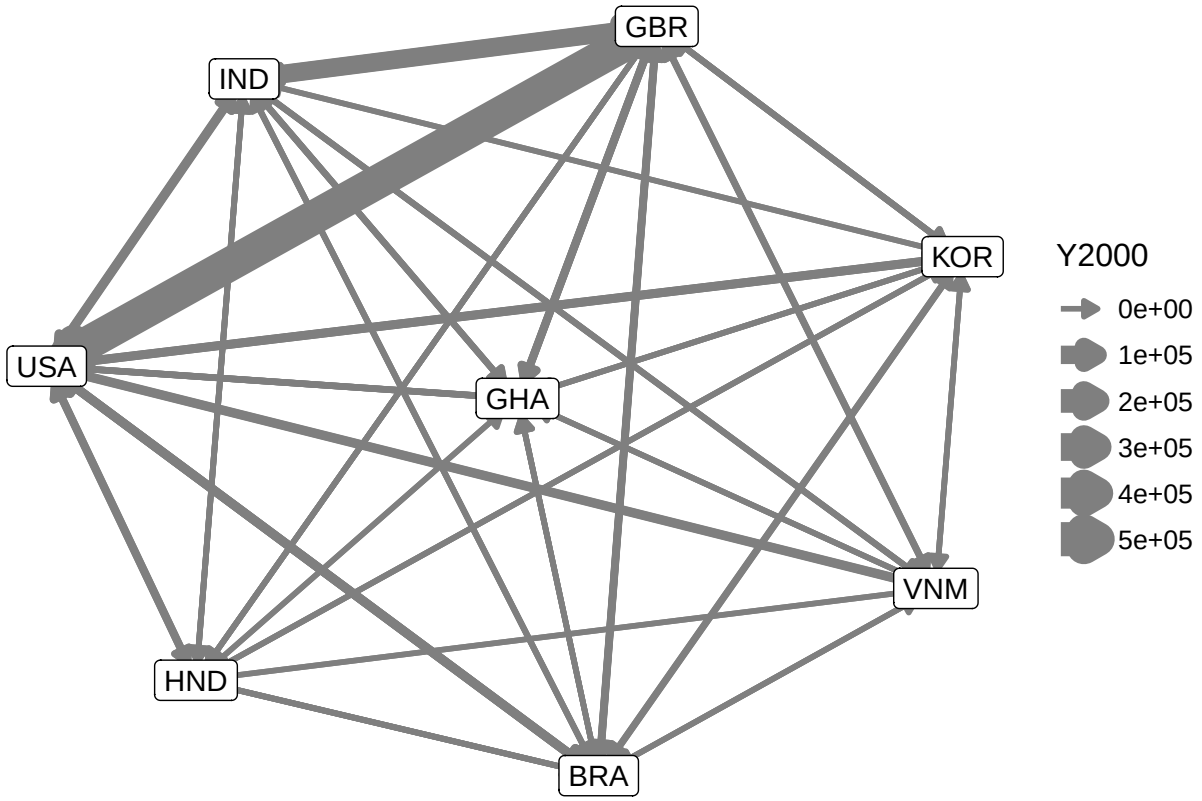
```
migration_network <- ggnetwork(migration)
```

```
## Warning in format_fortify(model = model, nodes = nodes, weights = "none", :
## duplicated edges detected
```

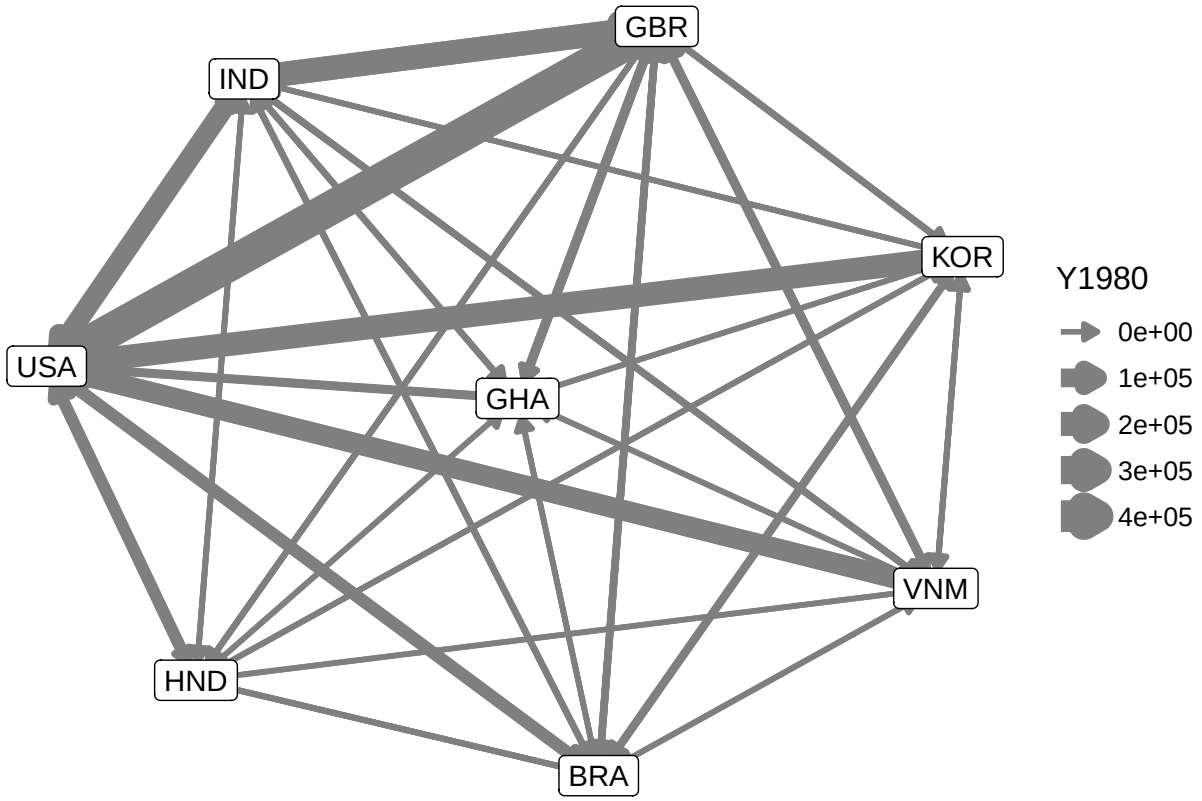
```
head(migration_network)
```

```
##   x         y name      xend      yend Y2000 Y1990 Y1980
## 2 0 0.5464625  USA 0.5841775 0.0167925  7025  6662  7126
## 3 0 0.5464625  USA 0.4891144 0.5055402   428    58    19
## 4 0 0.5464625  USA 0.6454376 0.9859298 66267 72740 64194
## 5 0 0.5464625  USA 0.1539348 0.1505646   958   422   547
## 6 0 0.5464625  USA 0.9469025 0.2571920     4    24    46
## 7 0 0.5464625  USA 0.9752631 0.6890789  3107  1132  1316
```

```
ggplot(data = migration_network
       , aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_edges(arrow=arrow(type="closed", length=unit(6,"pt"))
             , color = "gray50"
             , aes(size = Y2000)) +
  geom_nodes() +
  geom_nodelabel(aes(label = name)) +
  theme_blank()
```




```
ggplot(data = migration_network
       , aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_edges(arrow=arrow(type="closed", length=unit(6,"pt"))
             , color = "gray50"
             , 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: Brazil appears to have received immigrants from each of the seven countries and had an indegree of 16. After taking into account the distances from each other country, the value increases to 20,885.

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

```
## BRA GHA HND IND KOR GBR USA VNM
##  16  16  16  16  16  16  16  16
```

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

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

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: For the 1980 data, it appears as if the top five of origin are: Great Britain, India, South Korea, Vietnam, and the United States. Conversely, the top five of destination appear to be: the United States, Great Britain, Brazil, India, and South Korea. The top five countries of origin are the top countries from which people emigrate from and the top five countries of destination are the top countries to which people immigrate.

```
mig <- graph_from_data_frame(Migration)
strength(mig, weights = E(mig)$Y1980)
```

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

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

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

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

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

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: For the 2000 data, it appears as if the top five of origin are: Great Britain, India, the United States, South Korea, and Brazil. Conversely, the top five of destination appear to be: the United States, Great Britain, Brazil, India, and Ghana. The top five countries of origin are the top countries from which people emigrate from and the top five countries of destination are the top countries to which people immigrate.

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

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

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

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

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

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

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

> ANSWER: The diameter of the network is 527. Thus the two farthest nations are 527 components from each other.

```
diameter(mig, weights = E(mig)$Y2000)
```

```
## [1] 527
```

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

> ANSWER: The density of this network is 0.2857. This means that 28.6% of the connections that could exist, do exist.

```
V(mig)
```

```
## + 8/8 vertices, named, from 3b6609f:
## [1] BRA GHA HND IND KOR GBR USA VNM
```

```
amount = vcount(mig)
possible = vcount(mig)*(vcount(mig)-1)/2
amount/possible
```

```
## [1] 0.2857143
```

## 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: