STAT 231: Problem Set 7B

Evan Daisy

due by 10 PM on Friday, April 16

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

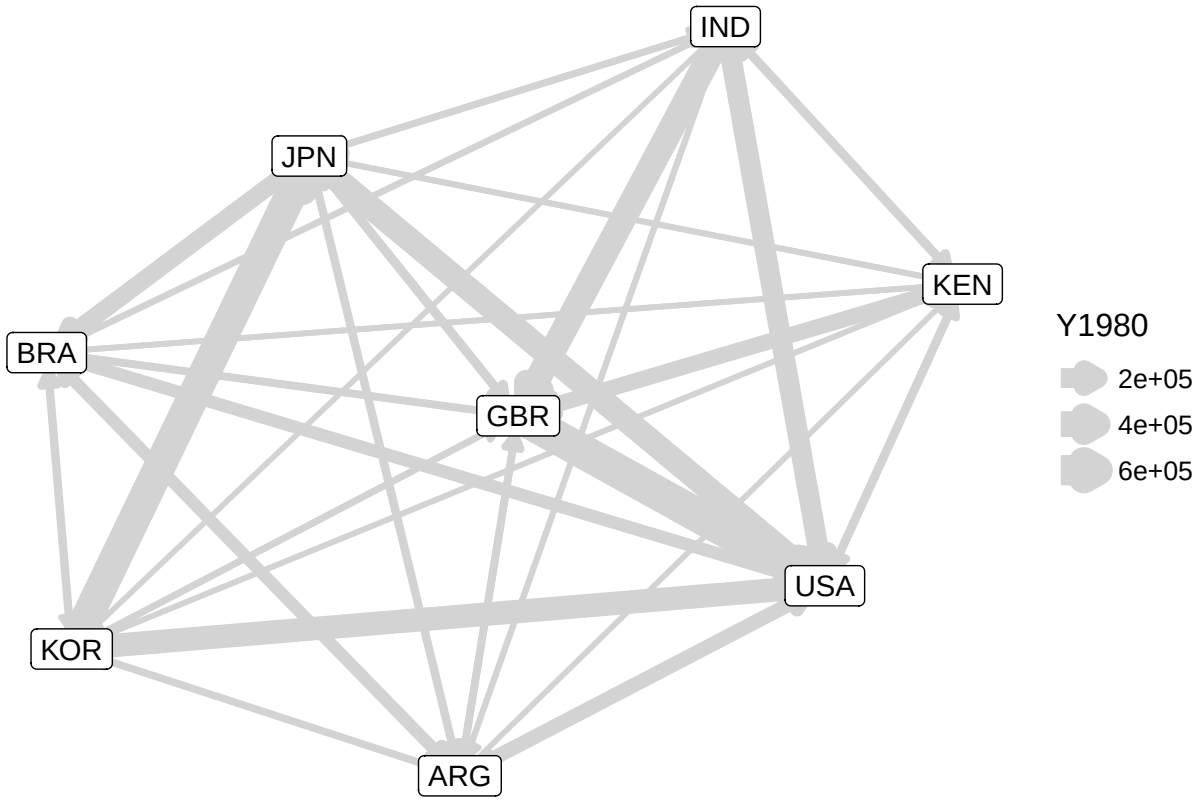
1a. Consider migration between the following countries: Argentina, Brazil, Japan, Kenya, Great Britain, India, South Korea, United States. Compare the TOTAL (males + females) migration between these countries over time. In separate (directed) graphs for 1980 and 2000, visualize the network for the these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

Don't forget to order the columns correctly and only keep relevant rows before transforming into a network object.

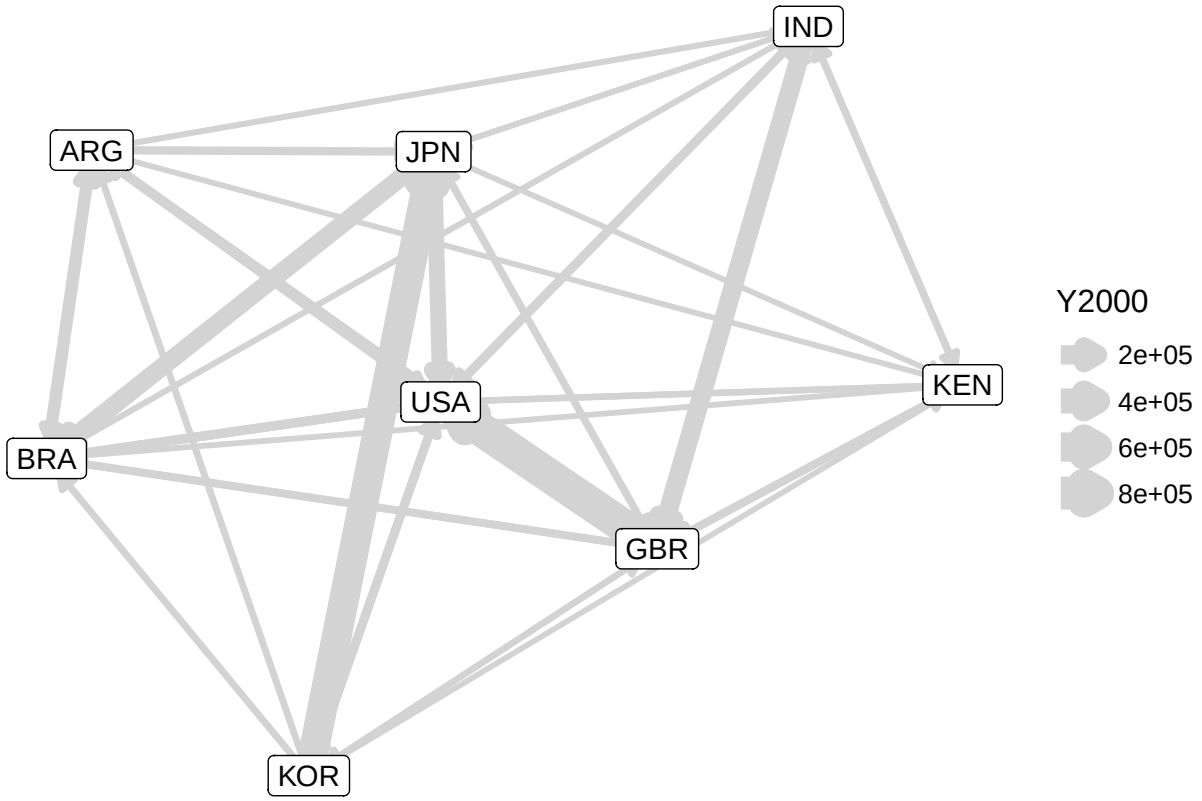
> ANSWER: The main message that I can see in these graphs is that migration between Asian countries (India, Korea, Japan) and the US was far more frequent in 1980 than it was in 2000.

```
path_in <- "/Users/evandaisy/Applications/Git/Data-science/Labs"
MigrationFlows <- read_csv(paste0(path_in,"/MigrationFlows.csv"))

# Argentina, Brazil, Great Britain, Japan, Kenya,  India, South Korea, United States
countries <- c("ARG", "BRA", "GBR", "JPN", "KEN", "IND","KOR", "USA")

# need migration overall:
# do some prelim data wrangling to combine numbers for males + females
migration_2000 <- MigrationFlows %>%
  filter(destcode %in% countries, origincode %in% countries) %>%
  group_by(destcode, origincode)%>%
  summarise(Y2000 = sum(Y2000)) %>%
  filter(Y2000 > 0)
migration_1980 <- MigrationFlows %>%
  filter(destcode %in% countries, origincode %in% countries) %>%
  group_by(destcode, origincode)%>%
  summarise(Y1980 = sum(Y1980)) %>%
  filter(Y1980 > 0)
mig_1980 <- migration_1980[, c(2,1,3)]
mig_2000 <- migration_2000[, c(2,1,3)]

#Create the network
migration2000 <- graph_from_edgelist(as.matrix(mig_2000[,1:2])
                                     , directed = TRUE)
migration2000 <- set_edge_attr(migration2000, "Y2000"
                               , value = migration_2000$Y2000)
migration1980 <- graph_from_edgelist(as.matrix(mig_1980[,1:2])
                                     , directed = TRUE)
migration1980 <- set_edge_attr(migration1980, "Y1980"
                               , value = migration_1980$Y1980)
migration_network <- ggnetwork(migration1980)
migration2_network <- ggnetwork(migration2000)

#Graph the network
ggplot(data = migration_network
       , 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()
```




```
ggplot(data = migration2_network
       , 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()
```



1b. Compute the *unweighted* in-degree for Japan in this network from 2000, and the *weighted* in-degree for Japan 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: There was migration from 6 of the 7 other countries considered to Japan in the year 2000. The total number of people migrating to Japan from these 6 other countries was 636946.

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

```
## BRA ARG GBR IND JPN KEN KOR USA
##   7   7   7   4   6   5   3   7
```

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

```
##     BRA     ARG     GBR     IND     JPN     KEN     KOR     USA
##  205254   73723  329949   23033  636946    1889    9437 1045722
```

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: Of these seven countries, the one from which the most people migrated to the others in 1980 is Korea (with 993074 total people migrating to the other seven countries) followed by Great Britain, India, Japan, and the United States. The one to which the most people migrated was the United States (with 1811537 total people immigrating from the other seven countries),

followed by Japan, Great Britain, Brazil, and Argentina. This could possibly be explained by the political instability in Korea in 1980, the economic success of the United States, and the proximity of Japan to Korea (as a place to escape political instability).

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

```
##     BRA     ARG     GBR     IND     JPN     KOR     USA     KEN
##  194211   69756  647261   22709  705948    8286 1811537   27898
```

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

```
##    BRA    ARG    GBR    IND    JPN    KOR    USA    KEN
## 105490 107802 832184 643587 502540 993074 180296 122633
```

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: Of these seven countries, the one from which the most people migrated to the others in 2000 is Great Britain (with 901279 total people migrating to the other seven countries) followed by Korea, Japan, India, and the United States. The one to which the most people migrated was the United States (with 1045722 total people immigrating from the other seven countries), followed by Japan, Great Britain, Brazil, and Argentina. This could be explained by the consistent economic appeal of the United States and the fact that Great Britain had seen better economic days.

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

```
##     BRA     ARG     GBR     IND     JPN     KEN     KOR     USA
##  205254   73723  329949   23033  636946    1889    9437 1045722
```

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

```
##    BRA    ARG    GBR    IND    JPN    KEN    KOR    USA
##  67542  41082 901279 206519 294863  11706 639215 163747
```

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

ANSWER: The diameter of this network was 2 in 2000, indicating that for a pair of countries A and B that did not have people migrating directly between them, there exists a country C connecting them so that people migrated from A to C and from C to B or vice versa. Thus the shortest path between two countries is 2.

```
diameter(migration2000)
```

```
## [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 was 0.82, indicating that 82% of the possible migratory directions between these 8 countries were migrated along in the year 2000.

```
ecount(migration2000)
```

```
## [1] 46
```

```
8*7
```

```
## [1] 56
```

```
46/56
```

```
## [1] 0.8214286
```

2. Mapping spatial data

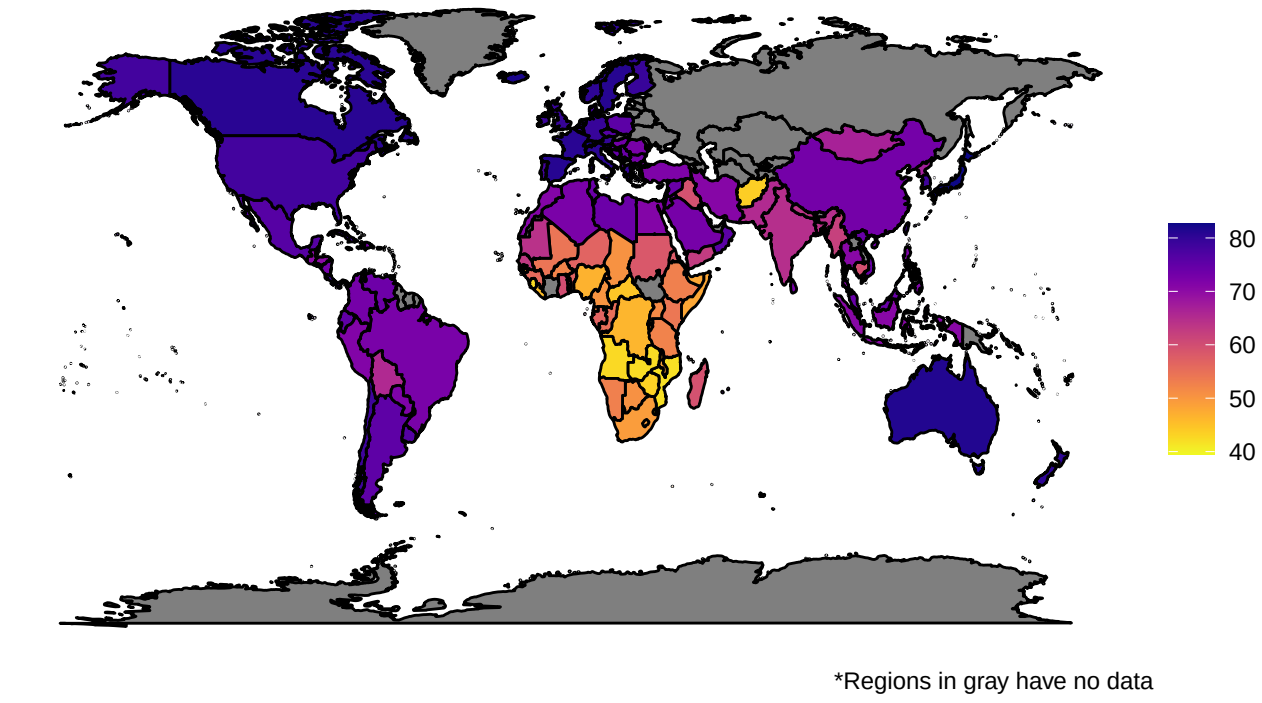
Reproduce the map you created for Lab08-spatial (and finish it if you didn't in class). In 2-4 sentences, interpret the visualization. What stands out as the central message?

NOTE: you do NOT need to say what colors are representing what feature (e.g, NOT: "In this map, I've colored the countries by GDP, with green representing low values and red representing high values") – this is obvious to the viewer, assuming there's an appropriate legend and title. Rather, what *information* do you extract from the visualization? (e.g., "From the choropleth below, we can see that the percent change in GDP per capita between 1957-2007 varies greatly across countries in Central America. In particular, Panama and Costa Rica stand out as having GDPs per capita that increased by over 200% across those 50 years. In contrast, Nicaragua's GDP per capita decreased by a small percentage during that same time span.")

> ANSWER: From the choropleth below, we see that there is little variation in life expectancy across countries in North America and Europe but that life expectancy varies widely across countries in Africa, with South America and Asia exhibiting intermediate variation (with the exception of Afghanistan). In particular, the band of countries including Angola, Zambia, Botswana, and Mozambique has astonishingly low life expectancy compared to the rest of the world.

```
# example of dataset with country information across years
library(gapminder)
gapminder <- gapminder::gapminder

# example of dataset with state information from 2013-2014
#library(fivethirtyeight)
#hate_crimes <- fivethirtyeight::hate_crimes

# example of another dataset with state information from 1977
# from the datasets package loaded above
#states_1977 <- as.data.frame(state.x77) %>%
#  add_rownames(var = "State") %>%
#  janitor::clean_names()

# example of dataset with county-level information from 2019
# see second tab in excel file for variable explanations
#county_employment <- readxl::read_excel(paste0(path_in, "/Unemployment.xls")
#                                        , sheet = 1
#                                        , skip = 7) %>%
#  janitor::clean_names()
recent_stuff <- gapminder %>%
  filter(year == 2007) %>%
  select(country, lifeExp) %>%
  rename(region = country) %>%
  mutate(region = case_when(region == "United States" ~ "USA",
                            TRUE ~ as.character(region))) %>%
  mutate(region = case_when(region == "United Kingdom" ~ "UK",
                            TRUE ~ as.character(region))) %>%
  mutate(region = case_when(region == "Congo, Dem. Rep." ~ "Democratic Republic of the Congo",
                            TRUE ~ as.character(region))) %>%
  mutate(region = case_when(region == "Congo, Rep." ~ "Republic of Congo",
                            TRUE ~ as.character(region))) %>%
  mutate(region = case_when(region == "Korea, Rep." ~ "South Korea",
                            TRUE ~ as.character(region))) %>%
```

```
  mutate(region = case_when(region == "Korea, Dem. Rep." ~ "North Korea",
                            TRUE ~ as.character(region))) %>%
  mutate(region = case_when(region == "Yemen, Rep." ~ "Yemen",
                            TRUE ~ as.character(region)))
world_map <- map_data(map = "world"
                      , region = ".")
lifeExp_map <- recent_stuff %>%
  right_join(world_map, by = "region")
ggplot(lifeExp_map, aes(x = long, y = lat, group = group
                      , fill = lifeExp)) +
  geom_polygon(color = "black") +
  theme_void() +
  coord_fixed(ratio = 1.3) +
  labs(title = "Life Expectancy By Country*"
       , subtitle = "as of 2007"
       , fill = ""
       , caption = "*Regions in gray have no data") +
  scale_fill_viridis(option = "plasma", direction = -1)
```



3. Mapping spatial data at a different level

Create a map at the world, country, or county level based on the choices provided in lab08-spatial, that is at a DIFFERENT level than the map you created for the lab (and included above). For instance, if you created a map of US counties for the lab, then choose a country or world map to create here.

Note: While I recommend using one of the datasets provided in the lab so you don't spend a lot of time searching for data, you are not strictly required to use one of those datasets.

Describe one challenge you encountered (if any) while creating this map.

> ANSWER: This to me seemed easier than the global map, especially since I didn't need to rename any regions. It was important to remember to make the state names all lowercase in the hate_crimes dataset before joining though, because otherwise they wouldn't have matched up.

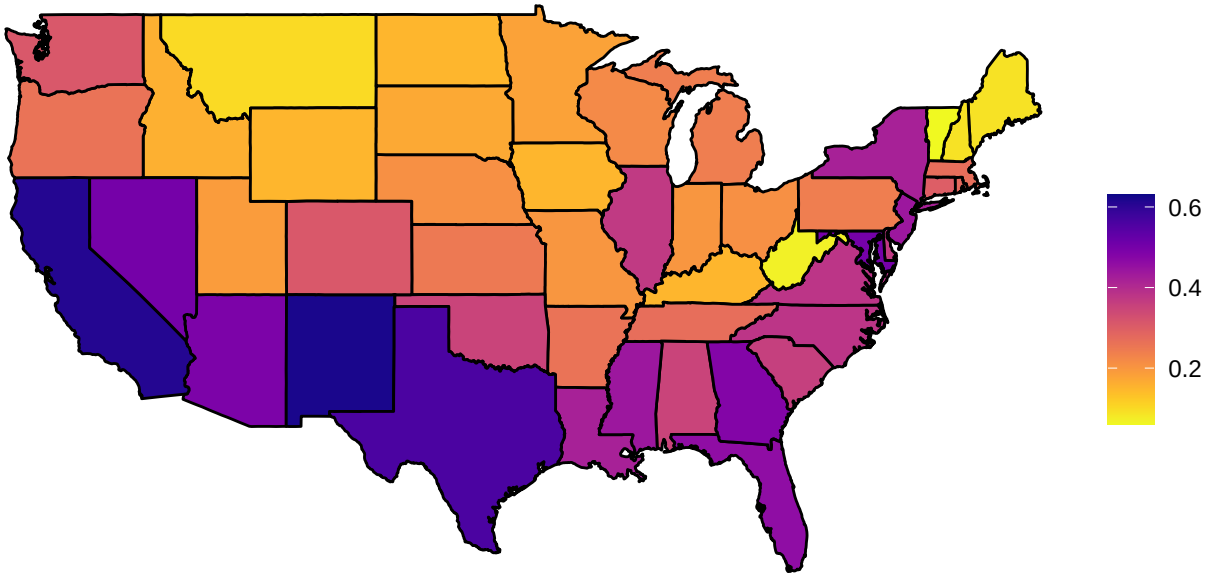
```
library(fivethirtyeight)
```

```
## Some larger datasets need to be installed separately, like senators and
## house_district_forecast. To install these, we recommend you install the
## fivethirtyeightdata package by running:
## install.packages('fivethirtyeightdata', repos =
## 'https://fivethirtyeightdata.github.io/drat/', type = 'source')
```

```
hate_crimes <- fivethirtyeight::hate_crimes
usa_states <- map_data(map = "state", region = ".")
non_white_map <- hate_crimes %>%
  select(state, share_non_white) %>%
  mutate(state = tolower(state)) %>%
  right_join(usa_states, by = c("state" = "region"))

ggplot(non_white_map, aes(x = long, y = lat, group = group
                      , fill = share_non_white)) +
  geom_polygon(color = "black") +
  theme_void() +
  coord_fixed(ratio = 1.3) +
  labs(title = "Proportion of Non-white Residents"
       , subtitle = "as of 2015"
       , fill = "") +
  scale_fill_viridis(option = "plasma", direction = -1)
```

Proportion of Non−white Residents

as of 2015



4. 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: The network is connected, because there is no set of two characters that don't have direct or indirect interactions connecting them (there is a path leading from any character to any other character).

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: Emma Thompson has an unweighted degree of 8, and Keira Knightley has an unweighted degree of 6. This means that Emma Thompson interacts with eight of the other main characters pictured, whereas Keira Knightley interacts with 6.

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: Hugh Grant would have a greater betweenness centrality. This indicates that Hugh Grant lies on more shortest paths between two characters, meaning that he serves as a more important "mutual friend" or at least mutual interactor in terms of connecting people.

5. Migration network on a world map! (OPTIONAL PRACTICE)

This problem is *optional* and will not be graded, but is given to provide additional coding practice and as a challenge to incorporate networks and mapping techniques together.

Create a world map that visualizes the network of countries we examined in #1 for the year 2000. For example, arrows to and from each of countries on the world map could have edge widths relative to their weighted degree centrality to represent migration to and from the countries.

Code to get you started is provided below.

```
# from mdsr package
# should see 'world_cities` df in your environment after running
data(world_cities)

# two-letter country codes
# Argentina, Brazil, Great Britain, Japan, Kenya
# India, South Korea, United States
countries2 <- data.frame(country3=countries
                         , country2 = c("AR", "BR", "GB", "JP"
                                        , "KE", "IN","KR", "US"))

# find capitals for anchoring points; can't find D.C., use Boston
cities <- c("Buenos Aires", "Brasilia", "London", "Tokyo", "Nairobi"
             , "New Delhi", "Seoul", "Boston")

anchors <- world_cities %>%
  right_join(countries2, by = c("country" = "country2")) %>%
  filter(name %in% cities) %>%
  select(name, country, country3, latitude, longitude)

# one suggested path:
# 1. based on the anchors dataset above and your Migration 2000 dataset created for # 1,
#    create dataframe that would supply geom_curve with the relevant arrow locations
#    (start points and end points)
# 2. create world map dataset using `map_data` function
# 3. use geom_polygon to create world map, geom_point and/or geom_text to add
#    city points, and geom_curve to add weighted/colored arrows
```