**LAB 4 – Flights and Network Traversal**

NAME \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

DUE \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The primary objective of this lab is to measure network distance on a global flight system, and to pick out central and peripheral airports. The secondary objective of this lab is to add new edges to see if the network can be improved, and to see how it is affected when we removed nodes. There are many amazing flight resources out there---including data on number of passengers (by airport, by month), ticket price, flight time, real-time airplane location, as well as interactive systems (<https://openflights.org/>).

**SECTION 1**

**Part 1: Analyzing the global flow system**

Step 1. Download data from Canvas. You will have an edgelist with coordinates for R called R\_Edgelist.csv. You will also have two files in the Gephi folder called *edgelist* and *nodelist*.

This data was downloaded from <https://github.com/jpatokal/openflights/tree/master/data>. We are evaluating the network structure for now, not the passengers.

IF YOU BRING YOUR EDGELIST INTO GEPHI, remember it is **undirected**. And click “**append to existing workspace**” if you bring it in after your nodes.

Step 2. Report on critical statistics of these flows using any software you’d like.

1. What is the airport with the highest degree? (for these questions, don’t just report the code, also say where it is) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (ask: why does Juneau only have 1?).

2. Which airport has the highest betweenness centrality (and what is it)? Comment on this result in the context of where it is in the world, its population, ‘power’, etc.. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

3. Which airport has the highest closeness centrality (and what is it)? Comment on this result in the context of where it is in the world, its population, ‘power’, etc.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

4. What is the network diameter (longest shortest path?) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_? What is the average network diameter (e.g. the average number of hops) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

5. How would you get the average diameter to *decrease*? That’s a good thing. Give a very clear strategy using 3 sentences, you will implement it below.

6. What is the *eccentricity* of Atlanta? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_How many airports does it connect to with this number of hops? Choose one and give the flight path(layovers) (ex. “Boston to Chicago to London to Detroit”) between Atlanta and this destination? What is kind of ‘weird’ about the farthest places to travel? And what’s going on here? \*\*See R CODE \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

7. Repeat question 6 (flight path/layovers), but for Atlanta’s farthest (in hops) destination that is NOT in the US.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

Bonus +3: List the path between the two most distant airports that are the most removed from each other? (\*\*See R Code) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Graphic 1: Make a map with the flows from the airports with degree = 3. If you are using Gephi, you can try the Geolayout.** See this: <https://gephi.wordpress.com/tag/geolayout/>

**R code for Part 1 Step 2:**

**#install packages**install.packages("igraph")  
library("igraph")

**#bring in fulledgelist.csv —that’s all!**

R\_Edgelist <- read.csv("C:/Users/Clio/Desktop/R/R\_Edgelist.csv")

**#subset the fulledgelist**

edgelist <- R\_Edgelist[,1:2]

View(edgelist)

**#create a graph from the edgelist**

g <- graph\_from\_data\_frame(edgelist, directed = FALSE)  
print(paste("We have" ,vcount(g), "nodes."))  
print(paste("We have" ,gsize(g), "edges."))

**#Compute diameter statistics**diameter(g)  
mean\_distance(g)

**#Compute node statistics**bet <- betweenness(g)  
clo <- closeness(g)  
ecc <- eccentricity(g)  
NodeFeatures <- data.frame(cbind(bet, clo, ecc))  
View(NodeFeatures)

**## Examine some relationships**

plot(NodeFeatures$clo, NodeFeatures$ecc, log = 'x', xlab = "Closeness Centrality", ylab = "Eccentricity", main = "Comparison of Airport Features")  
plot(NodeFeatures$clo, NodeFeatures$bet, xlab = "Closeness Centrality", ylab = "Betweenness Centrality", main = "Comparison of Airport Features")

**##You can used GGPLOT2, or original R to experiment with changing the axis, labels, etc.**

plot(NodeFeatures$clo~NodeFeatures$bet, xlab = "Betweenness Centrality", ylab = "Closeness Centrality", col="lightblue", xlim=c(5000,max(NodeFeatures$bet)), pch=19, cex=2)

text(NodeFeatures$clo~NodeFeatures$bet, labels=rownames(NodeFeatures),data=NodeFeatures, cex=0.5, font=2)

**###Interesting: if you add more peripheral nodes (i.e. use the larger flight data set), even with a degree of 1, they would increase the ‘central’ nodes’ betweenness centrality by a lot!!**

**##Now time for hops. This network is absolutely transitive.**

Dist <- distances(g, mode="all")

AtlantaHops <- Dist["Atlanta\_US",]

**##What intermediary airports exist between Atlanta and ‘a’ farthest airport?**

shortest\_paths(g, "Atlanta\_US", "[INSERT YOUR AIRPORT NAME HERE]")

**PART 2: Perturbing the global flow system.**

**Step 1.** **Now, we will experiment with adding “strategic” edges to the network to see how we can decrease the overall hops.** **Choose either A or B.** **This is question #8.**

1. Your “strategy” can be qualitative, you can look at the network and try to decide which places should be linked together, and do some trial and error. Report your 3 different qualitative strategies. Hints: In one trial, join places that could use some ‘help’ (they are both kind of low on their centrality), and then in another trial, try joining places that are already ‘powerful’. Do a few trials. Make a **table with at least 5 rows** that shows the added leg, and the decrease in “mean\_distance” (average diameter) (in hops). Show your added legs (it’s easy to do in Gephi, use the pencil tool).

*Put your {TABLE} and DISCUSSION and MAPS with ADDED LEGS HIGHLIGHTED here.*

1. Your strategy can be “brute force!” You can use a **for** **loop** and try a bunch of **random** edges. You can add many edges and see how things accumulate, or you can just do one at a time.

Make a plot that shows your many trials (of adding one, or a few, or cumulatively) and how much they decreased the “mean\_distance” (average diameter) (in hops).

*Put your {PLOT} and COMMENTED CODE and DISCUSSION here.*

**##R Code for Part 2 Step 1**

**##Here is some code to help! You will have to change it.**

**##make yourself a copy of the edgelist**edgelist\_2 <- edgelist

**##Give it a new edge**edgelist\_2[nrow(7749,1] <- "[YOUR NODE]"

edgelist\_2[nrow(7749,2] <- "[ YOUR NODE]"

**##compute the mean\_distance**  
g2 <- graph\_from\_data\_frame(edgelist\_2, directed = FALSE)  
mean\_distance(g2)

*More help:  
Do you want to increment a new row? try:* ***nrow(edgelist\_2)+1*** *Do you want to choose a random airport? Look at your last lab. “****Sample****” is a good keyword.*

**Step 2. Now, we will see how the removal of a node or nodes means an increase.**

8. What very sad and serious global health crisis is occurring right now? Have any airports been shut down because of it?

Report on 3 statistics: the **mean\_distance** if a VERY LARGE city in this country is removed from the network; the **mean\_distance** if a DISEASE EPICENTER city is removed, and the **mean\_distance** if airports from the WHOLE COUNTRY are removed.

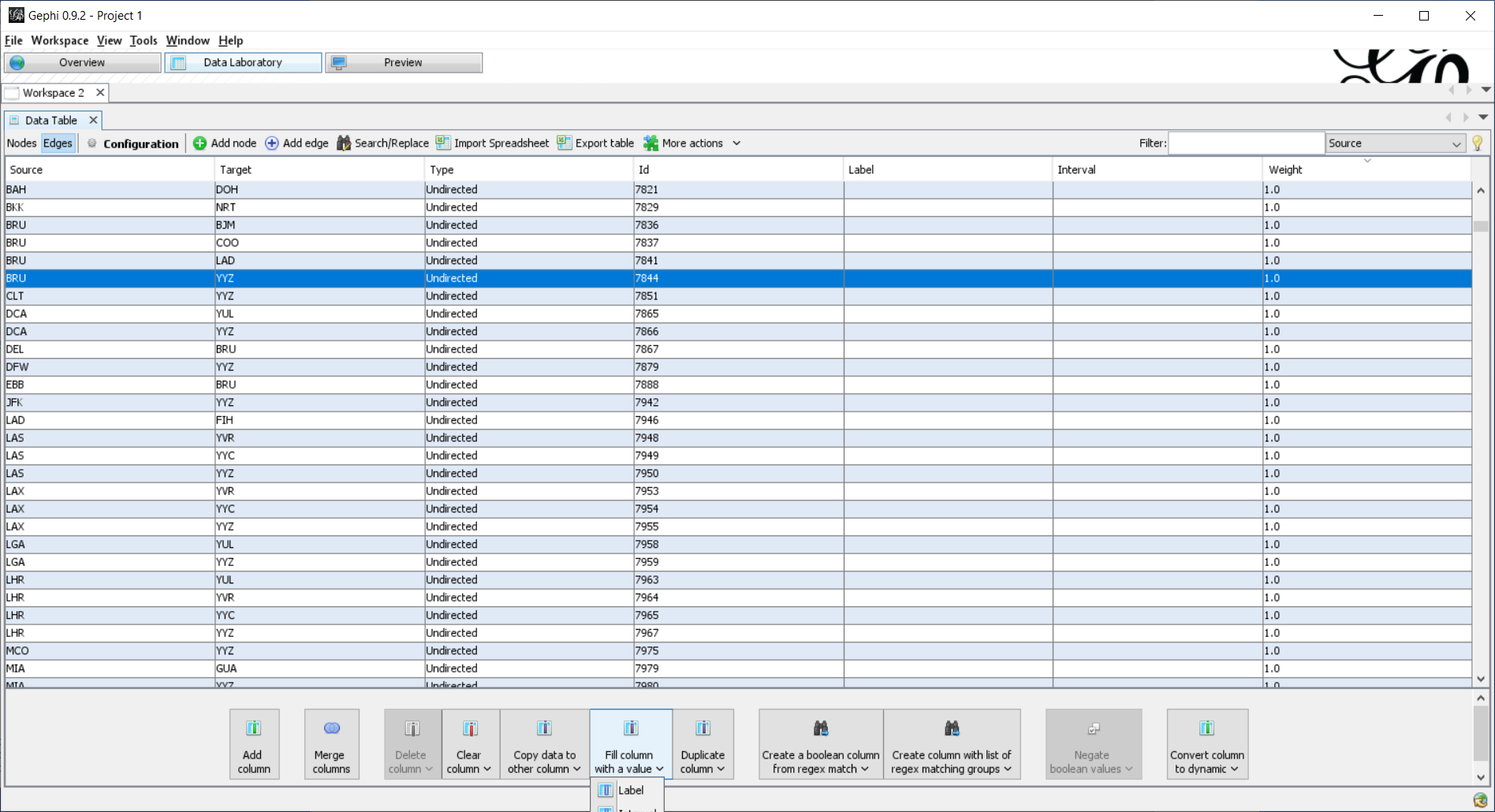
Why does taking out only 1 airport have a bigger effect than taking out a whole country?

**#Here is some code to help you create a new edgelist with removed nodes.**

edgelist\_3 <- sqldf("select \* from R\_Edgelist where target\_city NOT LIKE '%[NODE/KEYWORD]%' and source\_city NOT LIKE '%[NODE/KEYWORD]%'")

**TROUBLESHOOTER**

If your number of hops is greater than it seems like it should be, or it’s bigger than the diameter, go in to your edgelist, and click Fill column with value (select “1”). This should fix it.





Make sure this says 6, not something like 18. (This graphic may have more nodes than yours, that’s OK).

Can’t see your labels? Try “Label Adjust” in the Layout.

**SECTION 2**

Traveling in and out of Atlanta

Use the table called FlighTimes.xlsx that has the (actual) travel times in minutes for over 500,000 domestic U.S. Flights. These data are for the month of January, 2019. They are from the Bureau of Transportation Statistics, US. DOT <https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time>

Filter this dataset to only show flights that leave **ATL** **or** that come to **ATL**. Choose one option.

Do a SQL query to find the average time and standard deviation of the time of your chosen flights for the edges of your choice.

Example of what your data structure should look like if you are using ATL as the origin:

BEFORE 🡪   
origin, destination, time  
ATL, DCA, 88  
ATL, DCA, 80  
ATL, DCA, 95

ATL, DCA, 100

AFTER 🡪   
origin, destination, average time, standard deviation  
ATL, DCA, 85, 12

9. What is your most unreliable location? What were the min and max times and what was the standard deviation of times? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Join the non-ATL airports back to the coordinates you have for the airports and bring the points into your mapping software.

10. What is a problem with isochrones? How can we fix this? Do isochrones depend on mode? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**GRAPHIC** Make a nice map of the time it takes to get to these places (use color for the points).

a. Either use ‘fuzziness’ to illustrate the reliability of the time (watch my demo).

b. Or use contour lines to create ISOCHRONES.

c. Or do both!