

**Lab 3:**

**Heterogeneity and Friction (personal ties)**

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

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

We are mapping **dyadic** interactions. These edges form networks typically, but they can also be studied as standalone connections. It is quite rare to see them in GIS.

In Section 1, we will create critical statistics about these edges and their interactions over different boundaries and given different distances.

In Section 2, we compare random flows throughout the city to the dyads that were observed. They represent a ‘null model’ of how likely a relationship is to cross a certain boundary.

If *dyads* cross boundaries, more often than *random flows*, we consider the dyadic flows to be indicators of special relationships that might not occur by happenstance. In the figure above, there are four visual examples of how mentorship pairs (red) statistically differ the connection behavior of random pairs. 1a: pairs connect at a further distance than random pairs. 1b: pairs cross administrative school districts more frequently than random pairs. 1c: pairs connect areas with a distinguished difference in census values (delta = 30%, 40%) than random pairs (delta = 0%, 0%, 0%, 10%, 0%, 0%, 0%). 1d: pairs connect different commuter zones than where random connections travel.

The data in this lab come from a mentorship program at the Monte del Sol school in Santa Fe NM. See <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2811315>. It was extended for use with Big Brothers Big Sisters (paper in your extra readings folder on Canvas). **We are getting Atlanta data from BBBS soon!!!**

SantaFeBlockGroups are from the Census and variables include common demographics.

**SECTION 1**

**Part 1: Set up and edge review**

**Step 1:** Download all lab files from Canvas. Open ArcMap. Unzip this content and put it in ArcMap.

1. What do the points represent? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. What do the lines / edges represent? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. Open the Attribute Tables. What kinds of boundaries can edges cross? Give an example of each.
4. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
5. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
6. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
7. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
8. Theoretically, what does it mean when an edge crosses a boundary?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. What are different ways to measure dyadic edge **distance**?
2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
5. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
6. Theoretically, what does distance represent here?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Step 2: Change your data frame projection to NAD\_1983\_StatePlane\_New\_Mexico\_Central\_FIPS\_3002.** Is this a projected or geographic coordinate system?

Now, add a new field in Dyad Edges and call it “distance”. Make it a float. Calculate the distance in kilometers of these edges.

**Part 2: Likelihoods of Boundary Crossings**

**Step 1:** Do a **spatial join** with the dyad edges and *SantaFeHighSchools.shp*. Right click on DyadEdges to do so. Don’t change anything else. Output this file as ***Edges\_schools.shp***. **Check in: how will you know if the line crosses the high school boundary?**

**Step 2:** In the output file *Edges\_schools.shp*, make a new column called crossSchool. Make it a short integer. In the field calculator, put a 1 if the column crosses the boundary and a zero if it doesn’t.

1. What percentage of edges cross the boundary? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (Hint, highlight the number of rows that = 2 and see how many rows are selected, then divide that over the total number of rows.)

**Part 3: Allometric Distance**

**Step 1:** Allometric distance here is measured by the number of people who live in between the mentor and the protégé. This represents the number of “intervening opportunities”.

1. Why is this especially helpful in Santa Fe?

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Step 2:** Open the file called *SantaFeBlockGroups.shp*.

1. What field will you use to sum up population? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Which summary statistic will you use? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Step 3: Do a spatial join to attach the **SUM** of the block group population to the edge (you can use either ***Edges\_schools.shp*** or ***dyadEdges.shp***- it doesn’t matter). Output the file as ***Allometric.shp***.

*Allometric.shp* will have a lot of extra fields. **Check in: do these extra fields make *any* sense? Why or why not?** Hint: think of what it means to sum up percentages….

Step 4: Now make a scatterplot comparing allometric distance and Euclidean distance. Remember, your allometric distance is a measure of cost! How does this cost correlate with Euclidean distance (describe in the graphic caption below). **What a one benefit of creating a chart in ArcMap (or perhaps QGIS?)**

**GRAPHIC 1:** **Scatterplot** comparisonof Allometric Distance and Euclidean Distance. Put the scatterplot here and give it a caption. Explain why you see what you see. DO NOT INCLUDE A LEGEND. Explain what is going in on the image. Talk about in what cases allometric distance is useful and in what cases it isn’t useful. Remember to make nice axis labels and a nice title.

**Part 4: Socio-Economic Differences Part 1**

**Step 1:** **Open *dyadEdges.shp***. Choose a socio-economic variable that communicates that there are situations where mentors live in neighborhoods that are different than their protégés.

Make a map of how these socio-economic differences are manifested as flows. Hint: Create a ratio or difference of the same variable for both P and M, and color the edges based on this ratio or difference. Put your map below and describe what you find! Don’t forget your legend.

|  |
| --- |
| **GRAPHIC 2:** **Map** showing disparities for a socioeconomic indicator (ex. Difference in crimes or median HH income). Color your edges based on that indicator and color the underlying polygons similarly, but with a more transparent look. Explain what is going in on the image. Make sure your legend has nice round numbers as does your scale bar! |

**Step 2:** Open the attribute table of *dyadEdges* and export **dyadsedges** to a .txt. Remember to choose .txt as your output files when you export. You can call it dyadsedges.txt.

**Step 3: Open R Studio.** Import your .txt

10.Which variables need normalization? **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

11. What does normalization mean in this context? **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Step 4:** Use R studio to calculate the mean and standard deviation for HHincome for the protégés vs. mentors. The standard deviation tells us whether there is a wide range in those pools of people.

Type in commands like this into the console:

#comment: finds the mean of the of a value

mean(dyadEdges$p\_medHHinc)

#comment: finds the standard deviation for a value

sd(dyadEdges$p\_employed)

**Table of Means:**

|  |  |  |
| --- | --- | --- |
|  | **HHincome (mean)** | **HHincome (st deviation)** |
| **Mentors** |  |  |
| **Proteges** |  |  |
| **Difference in means** |  | N/A |

**Step 5**: Do a simple paired T test to see if the differences are statistically significant.

###This returns the t test statistic and p value that shows whether the average income of the proteges is significantly different than that of the mentors. P values less than 0.05 generally mean that there is a big difference in these distributions.

t.test(dyadEdges$p\_medHHinc, dyadEdges$m\_medHHinc, paired = TRUE)

**T Test Results:**

|  |  |
| --- | --- |
|  | **HHincome** |
| **T Test P Value and T statistic** |  |

**Step 6:** Make probability density function (PDF) distributions of the HH income distribution so we can see the difference. This means that you will create one plots of how collective protégé and mentors incomes are different.

//comment: This plots the density distribution of average household income of protégés – given the census block group they live in. the next line plots the second line on the same plot…in red! If you find that one line goes off the plot, switch the order in which you plot the variable. Here, I plotted protégés first, and mentors second. But I could have plotted mentors first and protégés second.

plot(density(dyadEdges$p\_medHHinc), main = "Great Title", xlab = "My X Variable")

lines(density(dyadEdges$m\_medHHinc), col = "red")

|  |
| --- |
| **GRAPHIC 3: Probability density distribution** TITLE. Explain what is going in on the image. |

**Part 5: Rewiring Edges**

**Step 1:** To make a “null hypothesis” probability density function (PDF) of HHincome, **we need to rewire the edges so that \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_,**

#how many iterations do you want?

n <- 100

#pull random sets of each edge and the accompanying HHincome # **Below: why 2,16 and 3,17?**

P <- dyadEdges[sample(nrow(dyadEdges), n, replace = TRUE), c(2, 16)]   
M <- dyadEdges[sample(nrow(dyadEdges), n, replace = TRUE), c(3, 17)]

#put them together as a dataframe

randomPairs <- data.frame(cbind(P, M))

#Plot the difference of randomPairs

plot(density(randomPairs$p\_medHHinc-randomPairs$m\_medHHinc), main = "Great Title", xlab = "My X Variable")   
lines(density(dyadEdges$p\_medHHinc-dyadEdges$m\_medHHinc), col = "red")

**12. What does the red line mean and what does the blue line mean? What does this plot show you? (No need to include the plot). But start your sentence with “the plot comparing \_\_\_\_ to \_\_\_\_\_ shows….”**

Rerun your T test for the randomly-assigned pairs.

t.test(dyadEdges$p\_medHHinc, dyadEdges$m\_medHHinc, paired = TRUE)

**T Test Results:**

|  |  |
| --- | --- |
| Random assignments | **HHincome** |
| **T Test P Value and T statistic** |  |

**Step 2:** What if anyone could be matched with anyone? This would be able to show if there is a difference between these two populations.

#rename your files just in case

P1 <- P  
M1 <- M

#rename column titles

names(P1) <- c("person", "HHinc")  
names(M1) <- c("person", "HHinc")

#combine the files into a long list

allPeople <- data.frame(rbind(P1,M1))

View(allPeople)

#randomly choose people to match together

First <- allPeople[sample(nrow(allPeople), n, replace = TRUE), ]  
Second <- allPeople[sample(nrow(allPeople), n, replace = TRUE), ]

#put them together as a dataframe

allPeopleRandom <- data.frame(cbind(First, Second))

#plot the difference of allPeopleRandom vs. Original Pairs

plot(density(allPeopleRandom[,2] - allPeopleRandom[,4]), main = "Great Title", xlab = "My X Variable")

lines(density(dyadEdges$p\_medHHinc- dyadEdges$m\_medHHinc), col = "red")

13. Can one person be “matched” with themselves here? Why or why not?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. What would happen if we said replace = False? Hint: Think of the value you gave ‘n’

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Step 3:** Turn your plots off by commenting them out. Run this code:

#means of real difference; random pairs; random pairs of anyone

c(mean(dyadEdges$p\_medHHinc-dyadEdges$m\_medHHinc), mean(randomPairs$p\_medHHinc - randomPairs$m\_medHHinc), mean(allPeopleRandom[,2] - allPeopleRandom[,4]))

1. What is dyadEdges? What is randomPairs? What is allPeopleRandom? (in your own words) and what is the mean of each (the output from the code right above)?

Which of the 3 variables retains the bi-partite nature of the network which removes the bi-partite rule?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. As you just wrote, the average mentor lives in a neighborhood that makes $2889.47 more. Is the difference in hh income for the actual pairs (dyadEdges) different than the randomPairs and allPeople random? How would you tell this to the program in a very explainable way? Hint: you can use a term like “your pool of mentors”.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Check in: How many times did we run our random simulation, was that enough?**

**BONUS:** 5 PTS: Rerun difference in HHincome for **randomPairs** and **allPeopleRandom** at least 10000 times. (use a for loop). Make a density distribution or perform an unpaired t test comparing the income differences in randomPairs to income differences for allPeopleRandom. What is the mean difference in 1000 iterations of random pairs, and what is the mean difference in 1000 iterations of allPeopleRandom?

**Part 6: Socio-Economic Differences Part 2, your turn!**

**Step 1:** The census block groups file you have in ArcMap is called SantaFeBlockGroups. Socioeconomic variables include: **u15pop, u18pop, pct65pl, medage, pcthisp, pcthsonly, pcteng**. Which stand for: percent aged under 15, percent aged under 18, percent over 65 years old, median age, percent Hispanic, percent with only high school degree, percent English speaking households.

Choose **one** of these variables and **repeat all of part 4 (not part 5) ☺** for that one variable. Hint: to have the biggest impact, use your choropleth map to look at which variable has a big difference between protégé locations and mentor locations.

You will have to use an **ANNOYING** combination of spatial joins TO ABSORB POYLGON INFORMATION TO YOUR POINTS (FIRST) and then TWO table joins (SECOND) TO ATTACH THE DATA TO THE EDGES.

Checklist for part 6:

**Image 1:** Map of symbolized flows. Use thickness this time.  
**Table** **1**: Table of means, difference in means, standard deviations, p value from t test for your variable.   
**Image 2:** Probability densify function (PDF) graph.  
**A ~100-150 word paragraph** describing exactly what you found out. Cite numbers and the map. Look up “Santa Fe” on Google Maps if you want to describe things better.

C:\dropbox\cma24\Dropbox\Friendly Cities\Xi_Cassie_Clio\LosAngeles_HowDoesThisMapLook.tif

Different mentor and protégé ties in Los Angeles shows clustering of protégés in the inner city, and clusters of mentors closer to the coast, and on the periphery. The edges make it hard to see, but the mentors typically live in higher-income neighborhoods.