Homework IV - Datascience

Nathan Bunch

Questions:

Give a definition of a graph or network.

A graph is a data type that shows a relationship between items. This is indicated by vertices labeled with various items and lines that connect them (also known as edges). These vertices and edges are represented by a visual, table, or matrix means.

Describe the three representations of graphs we looked at in class: visual, table, and adjacency matrix.

In terms of a graph being represented, a visual representation is by far the most common and noticeable. The visual representation has circles that represent each vertice and lines that depict each edge. In a table representation, it displays a list of items on one column and items in another column. This shows a relationship between one item and another (and therefore an edge). Finally there is the adjacency matrix. The adjacency matrix is a matrix whose labels for the rows and columns are the names of the items. If there is a relationship between one item and another, in the index represented by the row and column labels there exists a 1, if there is no relationship, there exists a 0.

Discuss some positive and negative things about the increase in connected online communities. What are some consequences of abstractly thinking about relationships between people as graphs?

The positive thing about having online communities is the ability to socialize with so many more people than one can originally without them. It connects people on opposite sides of the world or far away that would not have otherwise have met or been able to socialize. A detrimental thing about this is the influence of offensiveness. Because everyone can now be connected, if one were to accidentally send something that is offensive (or possibly offensive) then it can harm a larger number of people, rather than the small number that it would affect if the person was not connected. Now, a major consequence of abstractly thinking of

relationships between people as graphs is the lack of thinking of people as people, or possibly even the reasoning the connection exists. Once people become numbers and are objectified, then the feeling that a connection they share becomes almost dead, or not important as the feeling of that connection being a living, breathing thing doesn't exist anymore. We see this happening today. With some "online celebrities" having so many followers on major social media sites, the online celebrities don't care about their followers, because to them it's all "just numbers." These famous people sometimes fail to realize that they have an affect on their followers and that there is a legitimate connection between the celebrity and the follower online.

Real-world Data Project:

In this project, we were asked to take some data from everyday life and analyze it. The data had to come from something we come across on a daily basis. So, as part of that, I took the top three most basic questions every student comes across: "What is your name? What is your major? and What is your favorite color?" These questions come up both from a professor and/or other students. Anyway, the first step was to collect some data for analysis.

I started the process by creating a Google Form for students to enter some basic data about themselves. The form contained three simple questions:

What is your name? What is your major? and What is your favorite color?

The name and majors were both write-in options, as I wouldn't know all the student's names I was interviewing and also the majors of each student, however, with the colors….the options were limited to the ROYGBP color sequence (Red, Orange, Yellow, Green, Blue, and Purple). I was able to survey 92 different students with 52 different majors, some of which had double majors. There were a few entries that had been irrelevant data (like people putting in random text into the major and name sections and even just selecting a color and not placing a name and major).

When it came to cleaning the data, I separated the double-majors into their respective majors and kept the entries, even though it generated duplicate names. Because the analysis didn't rely on names, the duplicate name entries didn't matter, what mattered was the coordination with favorite color with a major. Also, during the cleaning process, I removed the extraneous

data from the dataset. After this was done, the CSV file was imported into R and then the names and timestamps were removed from the dataset.

Then the analysis process started.

During the analysis process, the duplicates in the dataset for majors were removed and then paired into an adjacency matrix with the colors they were associated with. After this, the matrix was converted into a form that allowed the graph function to create a graph based on the points. Then it was plotted using the plot function.

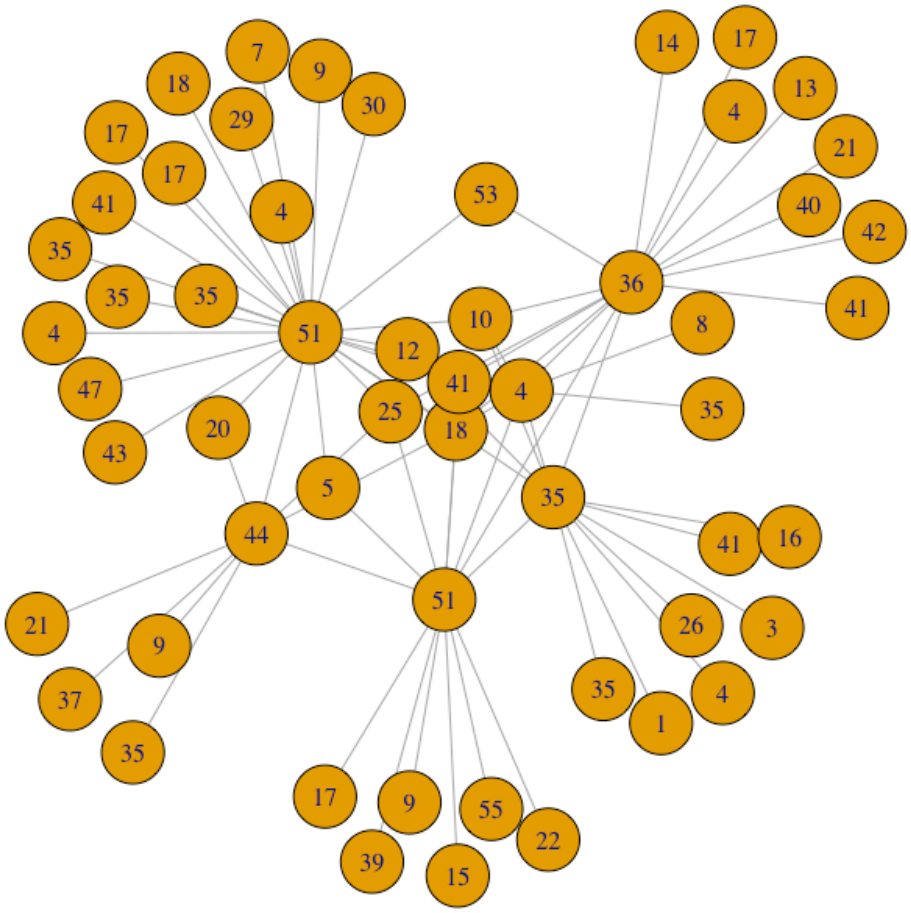
The computer-generated graph was good, but it wasn't exactly what I needed it to be, so I decided to finalize a better graph by hand. The computer generated graph can be found at the end of this document (the reason it didn't turn out well was due to the lack of better labels for the vertices). According to the computer-generated graph, there is an interesting phenomenon where there are small clusters of vertices. I believe this is where there are majors with only one color as their favorite. The other sections, particularly those in the middle, are where there are majors with two or more favorite colors in common. As should be noted is that there seems to be a ring around the inner vertices, these are the color vertices.

As you have most certainly noticed, it is very difficult to draw a conclusion from this computer-generated graph and therefore, that is why I had to make a hand-generated version of the graph, since I was unable to create labels on the graph using the computer.

Story continues…

So, after realizing that all I had done was incorrect…the graph and thew whole idea that I had in mind, I went back to square one and re-did my analysis…this time very frustrated. Instead of making the majors all vertices connected to other color vertices, I made the majors connected to each other by colored vertices. Because I could clearly not get this to work in R (R was being annoying again…after four hours of trying), I created the graph by hand. The finalized graph is in full color and shows all the major vertices connected by the colors of which they share in common. After that, I was finally relieved to know the assignment was finished.

Also, as a finishing touch, because some of the majors had more entries, I made those vertices larger than other vertices, to indicate the influence that they had on the overall percentage of favorite colors on campus.


7
9
18
30
29
17
17
41
4
35
35
35
4
51
47
20
43
53
10
12
41
4
25
18
5
44
51
35
14
17
13
4
21
40
42
36
41
8
35
41
16
26
3
35
4
1
21
9
37
35
17
9
55
22
39
15