Holly Gustavsen

I used a dataset I found on Kaggle called Music artists popularity. From the dataset, I took samples of size 100 and 200 and used the artist name (artist_lastfm) and their associated tags (tags_lastfm). I created an undirected graph where each artist was a node and an edge was created between two artists if they had a tag in common. I choose this dataset because I love listening to music and as a user of Last.fm have noticed that the tag system could be improved. There are a lot of redundant tags, for example the artist Rihanna has the tags rnb, r&b, and r'n'b. There are also a lot of hyper specific tags that are probably not useful such as many artists having their own name as a tag.

Originally, I was hoping to create a basic music recommendation algorithm which took a list of artists as the input and returned an artist that was not contained in that list and was most connected to the artists in that list. However, I realized this was too large of a project to complete in the time frame because with my chosen dataset, I had no way of testing if the recommended artist would actually be liked. I would have had to create my own secondary dataset to test the algorithm. So I decided to shift my project and study the degrees of separation between pairs of artists in the graph.

I created a function for small graphs (~5 artists) to calculate the actual average distance between all possible pairs of nodes in the graph. I ran this function a few times and the output was normally that the average distance was either 1 or slightly above 1, meaning almost every artist was connected to every other artist. An example of this output is below. This answer is very different from a social circle graph for example where it is expected that every person is connected within a distance of 6.

```
hollygust@crc-dot1x-nat-10-239-16-243 finalproj % cargo run
    Finished dev [unoptimized + debuginfo] target(s) in 0.11s
     Running `target/debug/finalproj`
There are 4 artists in this sample.
The artists in this sample are ["Coldplay", "Radiohead", "Red Hot Chili Peppers"
, "Kabylifornie"].
The average distance between artists in this sample is 1.0
The distances between each possible pair are [1, 1, 1, 1, 1, 1]
0 pairs of artists cannot be connected.
The pairs that can't be connected are [].
```

For graphs larger than 5, I calculated the average distance of n random pairs from various sample sizes. I tested this on sample sizes of ~10, ~20, and ~100 with 10, 100, and 1000 random pairs. While the average distance remained around 1, I was surprised that increasing the sample size did not cause the number of isolated pairs to reduce. The percentage of isolated pairs to the number of iterations is about 50% for all sample sizes and numbers of random pairs.

	~10 artists	~20 artists	~100
10 pairs	Ave. distance: 1 | # unconnected pairs: ~4	Ave. distance: 1.25 | # isolated: 6	Ave. distance: 1.2 | # isolated: 5
100 pairs	Ave. distance: 1 | # isolated: ~52	Ave. distance: 1.04 | # isolated: 43	Ave. distance: 1.14 | # isolated: 49
1000 pairs	Ave. distance: 1.17 | # isolated: 564	Ave. distance: 1.18 | # isolated: 528	Ave. distance: 1.15 | # isolated: 451

Based on these results, the usual distance between pairs of vertices for my graph (~ 1 degree) seems to be significantly lower than the expected distance of a social circle graph which is probably around 3. However, a social circle graph is thought to be fully connected within six nodes or degrees of separation. In contrast, my graph has some pairs that simply cannot be connected. In general, this makes sense because music can be categorized into genres that do not have a lot of overlap. However, based on the number of tags in the data, I was not expecting to have so many unconnected pairs. I assumed that in most cases there would be an artist who had both tags of two other artists and could connect them. For example, artist A has the tag “pop”, artist B has the tag “country”, and artist C has the tags “pop” and “country”. I believe that the unconnected pairs are the result of artists from very different genres such as classical and hip-hop, where even with Last.fm’s excessive tagging system, a connection would be rare and very difficult to produce. I believe if I were to use a significantly larger sample size or the whole dataset (1.4 million artists) the number of unconnected pairs would decrease or even reach 0 and the average distance between two nodes would increase.