Devin Rauscher

Professor Leonidas Kontothanassis

CDS DS 210 Programming for Data Science

7 May 2023

DS210 Final Project

Determining if the 6 Degrees of Separation Theory Works in Selected Boston Marathon Runners

For my final project, I decided to focus on data related to the Boston Marathon, mostly because the runners and finishers seemed interesting to examine and because I had the privilege of being able to watch the marathon for the first time this year. Although this project handles very limited data, I thought that it would be interesting to see if the theory of six degrees of separation held up around the marathon.

Data Selected

For my project, I focused first on a dataset that named the male and female winners of each Boston Marathon, from 1897-2022 for men and 1966-2022 for women. There was no marathon in 1918, due to U.S. involvement in WWI, and in 2020, due to the COVID-19 pandemic.

Because this dataset was not particularly large, at less than 200 points, I also included more data on the Boston Marathon; namely, all finishers in the 2015, 2016, and 2017 races. For these years, I mixed all years together and selected 1000 runners randomly.

The four columns of consideration, across all data, was year, name (of the runner), country of origin, and whether or not they won. For connections, runners were considered to ‘know’ each other if they came from the same country or ran in the same year. In this way,

degrees of separation were more related to similarities than personal knowledge, as someone who ran in 1890 would not actually know someone who ran in 2010. For the sake of this project, separation was more related to tangible connection.

Algorithm Implemented

I used breadth-first search (bfs) to draw connections between the runners, as suggested in the Piazza post. This was preferred over depth-first search (dfs) because I was testing not only connections, but specifically whether or not those connections would occur within six nodes.

Results

Interestingly, the theory of six degrees of separation was disproven first because every runner could not be connected to every other runner in some way, through any path: “There is no path of separation between Runner1 and Runner2.” Upon further investigation, it seems like that was mostly because some countries represented by the winners were not represented by the finishers from 2015-2017.

For this to occur, I hypothesize that 1000 runners randomly selected from 1966-2022 would have a better chance of being completely connected. This is because outside factors, like economic and social factors, may keep runners from a certain country from being able to participate in the Boston Marathon.