KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Mapping Post-Secondary Transfer Rates

Nathaniel Jones

Faculty Advisor: Dr. Andrew Wilson

INTRODUCTION

In recent years, there has been a *growing interest* in understanding the factors that influence students' choices in *selecting* and *transferring* between post-secondary institutions. One such factor is *geographical distance*, which has been highlighted by the American Council on Education's 2016 study. This study revealed that *57%* of incoming freshmen attending public four-year colleges enrolled within *100 miles* of their permanent residence, suggesting that proximity plays a significant role in students' college decisions. Furthermore, the same study proposed that some students may strategically use these nearby schools as *"steppingstones"* to transfer to larger institutions later in their academic journey. Considering these findings, this research aims to explore the relationship between geographical distance and post-secondary transfer rates.

DATA INTRODUCTION: The *2020-'21 CollegeScorecard* includes data for *6,654* post-secondary institutions for the following parameters:

- *INSTNM:* The Name of the School.
- *LATITUDE:* The North-South position of the School.
- *LONGITUDE:* The East-West position of the School.
- *OMENRAP_ALL:* The proportion of students that withdrew from the institution they originally started at and enrolled in another institution within 8-years.
- *UGDS:* The number of undergraduate/degree-seeking students that started in the Fall of 2020.
- *CONTROL:* Indicates whether a post-secondary institution is Publicly or Privately owned.
- *ICLEVEL:* Indicates whether a post-secondary institution is a four-year school or a two-year school.
- *CURROPER:* Indicates if a school is currently operating or not.
- *DISTANCEONLY:* Indicates if a school only offers online classes or not.

DATA WRANGLE: *Public four-year* post-secondary institutions that are *currently operating* and *do not primarily operate online* were used in the analysis of this project. Schools missing values for either their *latitude, longitude, or transfer rate* were dropped prior to analysis. After filtering the data, *1,544* schools remained.

METHODS

- **DISTANCE BETWEEN SCHOOLS:** The *latitude* and *longitude* of each school was used to find the *distance* between schools using *'the great-circle distance'* function which computes the shortest distance between two schools on the surface of a sphere.
- **NUMBER OF SCHOOLS WITHIN 100 MILES:** After finding the distance between schools, we identified the schools within *100 miles* of each other. This group represents a *neighborhood* of schools, giving students various *'options'* when considering a transfer.
- **CLUSTERING COEFFICIENT:** The *clustering coefficient* of a school is a measure that aids in understanding the degree to which schools are *interconnected*. This measure represents the likelihood that *neighboring schools are also neighbors* with each other. To calculate the clustering coefficient for a school, determine the number of *actual connections* between its neighbors and the *maximum possible connections* between them. The clustering coefficient is then calculated as the ratio of the actual connections to the maximum possible connections.
- **NETWORK TRANSITIVITY:** The *network transitivity* is a global view of the clustering coefficient and measures proportion of the *interconnected* schools within a network. It reveals the tendency of schools to form *clusters* or *tight-knit groups* based on their relationships with one another. For this project, the network transitivity was *74%* which suggests that *schools are more likely to be part of tight-knit clusters*, where the neighbors of a school are also neighboring.
- **GEOPANDAS:** GeoPandas is a powerful and user-friendly library for *handling*, *analyzing*, and *visualizing* geospatial data by extending Pandas' functionality with geospatial capabilities. This library, alongside *networkx* and custom user-defined functions were used to create the network map displayed in *figure 1*.
- **NETWORKx:** *NetworkX* is an open-source Python library designed for the *creation*, *manipulation*, *analysis*, and *visualization* of complex networks or graphs. It provides a wide range of tools and algorithms that can be applied to various types of networks, such as social networks, biological networks, transportation networks, and many others.
- **GGPLOT2:** Ggplot is a popular and powerful data visualization library for the R programming language. It is based on the *"Grammar of Graphics"* concept which aid in creating consistent and expressive visualizations.

RESULTS

Figure 1: University Transfer Rates between Schools within 100 Miles

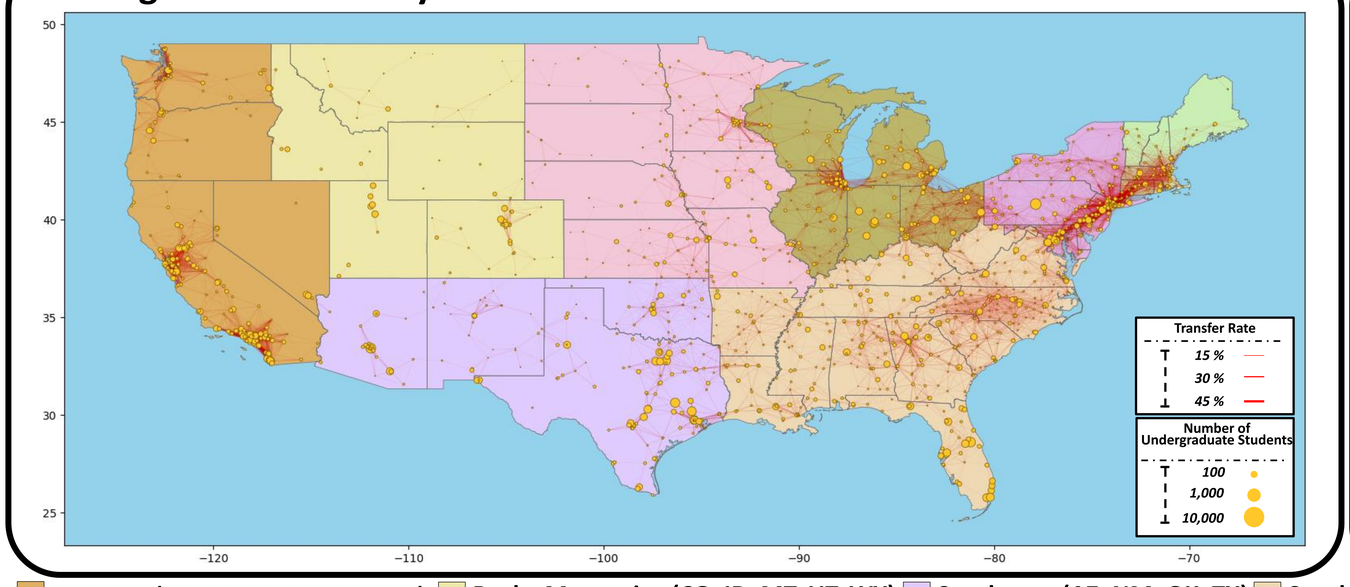
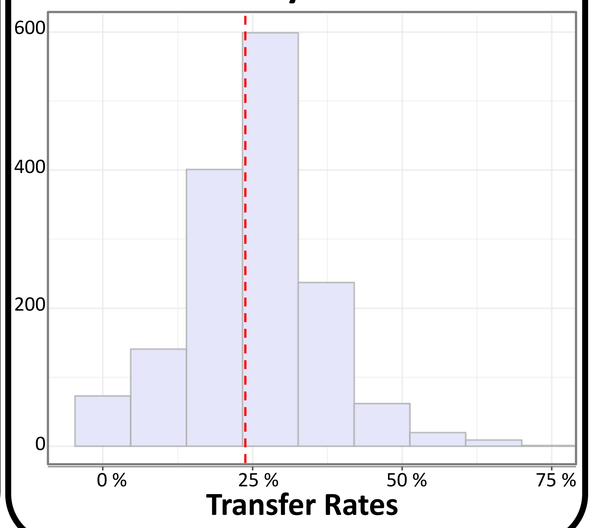


Figure 2: Histogram of Post-Secondary Transfer Rates



Far West (AK, CA, HI, NV, OR, WA) **Rocky Mountains (CO, ID, MT, UT, WY)** **Southwest (AZ, NM, OK, TX)** **Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV)**
Plains (IA, KS, MN, MO, NE, ND, SD) **Great Lakes (IL, IN, MI, OH, WI)** **New England (CT, ME, MA, NH, RI, VT)** **Mid East (DE, DC, MD, NJ, NY, PA)**

Figure 5: Number of Schools within 100 Miles by Region

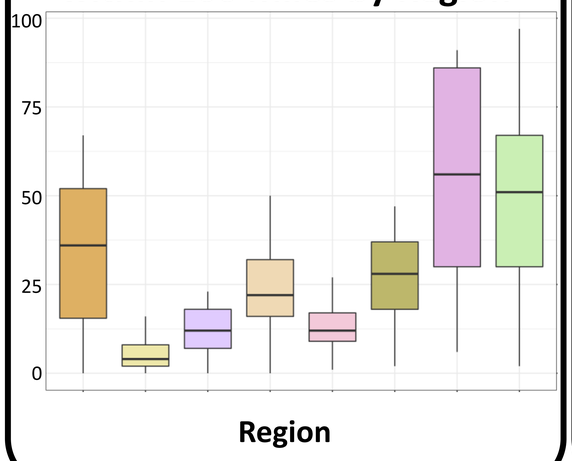


Figure 4: Number of Schools within 100 Miles by Transfer Rate

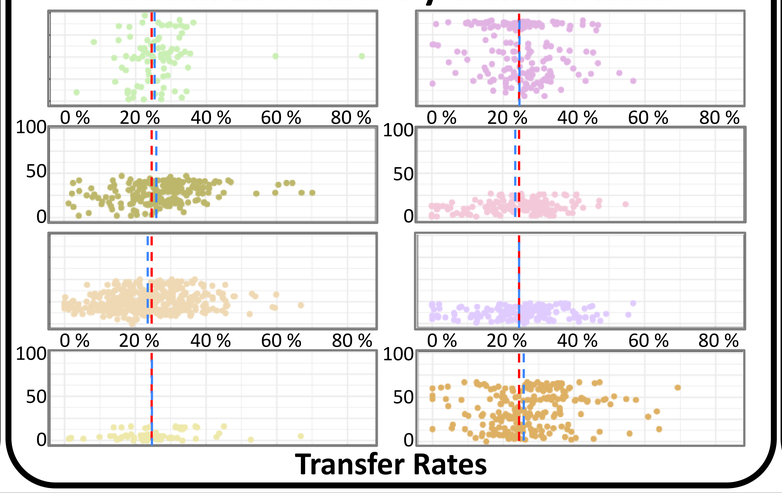
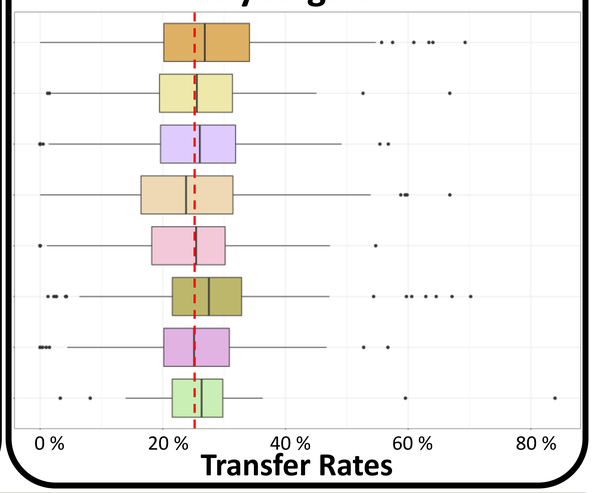


Figure 3: Average Transfer Rates by Region



CONCLUSIONS

- **University Transfer Rates between Schools within 100 miles (Figure 1):** Figure 1 shows the location of each school as a gold dot (or vertex) on a map. The size of each dot represents the number of undergraduate students at the school. When two schools are within 100 miles of each other, a red line (or edge) connects them. The width of this line reflects the school's transfer rate, with wider lines indicating higher transfer rates. From this visualization, we can observe that areas with fewer schools seem to have fewer and thinner connecting lines, which might suggest lower transfer rates.
- **Histogram of Post-Secondary Transfer Rates (Figure 2):** Figure 2 presents the distribution of transfer rates among post-secondary institutions. The average transfer rate, marked by a dotted red line, is 25%. We can see that most schools have transfer rates below 50%, indicating that less than half of the students transfer between institutions. The maximum transfer rate observed in the dataset is notably high at 84%. This suggests that there is a wide range of transfer rates among schools, and certain institutions may have unique factors or circumstances contributing to such high transfer rates. Considering the wide range of transfer rates observed in Figure 2, further analysis was conducted to explore the variation in transfer rates among different schools and regions.
- **Boxplot of Transfer Rates by Region (Figure 3):** Figure 3 displays the distribution of post-secondary transfer rates across different regions of the U.S. This plot shows the average transfer rate of each region does not significantly deviate from the overall average transfer rate observed in Figure 2 (visualized by the dotted red line). The region with the lowest average transfer rate is the Southeast, while the highest average transfer rate was observed by the Great Lakes region. Notably, although the average transfer rate of the Southeast region is less than the average transfer rate of the Great Lakes, the variability of transfer rates observed for the Southeast is much larger than the variability observed for the Great Lakes.
- **Scatterplot of the Number of Schools within 100miles by Transfer Rate (Figure 4):** Figure 4 displays the number of schools within 100 miles of a school on the vertical axis and the transfer rate horizontally. For each panel, the overall average post-secondary transfer rate is plotted as a dotted red line while the region-specific average is plotted as dotted blue line. Although some regions (Great Lakes) display a slight positive relationship between the number of schools within 100 miles and post-secondary transfer rate, this plot suggests the relationship may be very weak or nonexistent.
- **Boxplot of the Number of Schools within 100 miles by Region (Figure 5):** Figure 5 displays the distribution of the number of schools within 100 miles across different regions. This plot shows for each region the average number of schools within 100 miles, as well as the range of values observed for schools in that region. The Rocky Mountain region was observed with the lowest average number of schools within 100 miles. This suggests the Rocky Mountain region is very sparse in post-secondary institutions (which can be observed in figure 1). In contrast, the Mid East Region was observed with the highest average suggesting the post-secondary density of this region is very high.