An Alternative to Rank-Based College Exploration

Reed College

EJ Arce, Simon Couch, & Alexander Moore

May 2019

Table of Contents

Abstract

Rank-based college exploration, such as that popularized by U.S. News College Rankings, plays a pivotal role in how prospective college students discover and compare potential schools. However, despite its popularity, embedding hierarchy into college exploration reaffirms entrenched status obsession more than it educates and informs. We propose and implement an alternative method for prospective students and institutional researchers to evaluate colleges focusing on what makes a college similar to, not better than, another college. Data from IPEDS and the Mobility Report Card from the Equality of Opportunity Project are joined and provided to users in an R Shiny application that allows users to compare United States educational institutions through visualization, comparison of raw data, and model-based assessment of school similarity.

Chapter 1

Introduction

In March 2019, federal prosecutors charged over 50 celebrities, business leaders, and other wealthy people with participating in a bribery scheme to get their children into elite colleges. Among other things, parents paid to have test scores inflated, falsify biographies, and undergo bogus athletic recruitments (Medina, Benner, & Taylor, 2019). Since then, at the time of writing, there have been over sixty articles tagged with the keyword *college admissions scandal* in the New York Times alone ("College Admissions Scandal," 2019). Many of these articles have reflected on the cultural forces that could have contributed to such egregious and extreme actions. Three days after the news broke, Anemona Hartocollis wrote in the New York Times that "[a]t the heart of the scandal is a persistent adulation of highly selective universities. Elite colleges have become a status symbol with the legitimacy of meritocracy attached to them" (Stephens & Collins, 2019). Three days later, Bret Stephens wrote "... we've become a society that increasingly has a hard time distinguishing between the substance of a serious university education and the supposed benefits of a prestigious brand" (Hartocollis, 2019). The question, then, centers around the contributors to this status obession. In fact, this is a conversation that has been going on for decades: "[t]he students who are using [newsmagazine college] rankings are precisely those students... who already know, and act on, notions of which institutions are best. Newsmagazine rankings are merely reinforcing and legitimizing these students' status obsessions (McDonough, Lising, Walpole, & Perez, 1998)." This is one of many reasons rank-based college comparison models are fundamentally flawed and students deserve an alternative that provides more complete information for a significant life decision. Using holistic data on more than a thousand schools in the United States, this study proposes an alternative model that emphasizes similarities, rather than hierarchies, among college options.

Prospective undergraduates are not the only ones who confront the objective of evaluating and comparing colleges. Current college students who wish to transfer may seek another school similar to theirs in some ways but not others. Institutional researchers employed by a college are interested in identifying which colleges are similar to their institution to learn new practices that can foster better educational environments. Altogether, the shortcomings of current rankings systems affects many different individuals in academia.

Institutional characteristics such as academic expectations, cost, and location certainly factor into college decisions, but students typically do not have an idea of how exactly they want to use those characteristics to determine which college is best for them. This makes comparing colleges difficult when students have narrowed their list down to a handful of institutions and must pick a single school to attend. In this case, students (and college officials alike) would benefit from better methods of evaluation that would help to consider multiple factors about multiple colleges without spending an unrealistic amount of time researching every college individually.

Online services such as U.S. News attempt to ease this process of comparing schools by building models constructing a generalizable ranking system to evaluate which schools are best. Their models subjectively weight a number of quantitative measures that each college submits via survey. In addition to arbitrarily weighting college characteristics, U.S. News changes the weights in their model every year, presumably for media coverage (McDonough et al., 1998; Luca & Smith, 2013). The weighting for their 2018-2019 model is broken down below ("How U.S. News Calculated the 2019 Best Colleges Rankings," 2019):

- 35%: Outcomes (social mobility, graduation and retention, graduation rate performance)
- 20%: Faculty Resources (class size, faculty salary, and others)
- 20%: Expert Opinion (peer assessment, high school counselor assessment)
- 10%: Financial Resources (per student spending)
- 10%: Student Excellence (standardized tests, high school standing)
- 5%: Alumni Giving (percentage of alumni who donate to their school)

Multiple problems arise in response to algorithmic rankings. For one, college administrators embellish their numbers to look better than they actually are in hopes of receiving a higher ranking (Stecklow, 1995) (Kim, 2018). U.S. News editors have admitted to changing their model "on an annual basis in order to mollify their college critics" (McDonough et al., 1998). When Reed College, unsatisfied with these ranking procedures, refused to participate, U.S. News responded by deiliberately filling in missing information with unimpressive estimates, resulting in lower rankings (Lydgate, 2018). The placement of one school that embellished their submitted statistics over another school who provided honest statistics can attract a larger set of prospective students to the former instead of the latter, a problem that becomes important when students have narrowed their options to a handful of colleges and use rankings sites to inform final decisions, or even worse (and more common), when students do not consider schools below some arbitrarily chosen ranking (Ortagus, 2016; Furukawa, 2011).

There are several setbacks with the U.S. News (and more generally, hierarchical) model that perpetuates the problems described. Principally, these models seek to develop a universal measure based on arbitrarily chosen school characteristics and model weights. Shifting the goal of college comparison away from hierarchical methods towards one that compares how similar a number of colleges are to one another will provide a more honest and healthy method of evaluating colleges so students can make a more informed decision on where to pursue their studies. As a result, shifting away

from hierarchical comparison disincentivizes the embellishment of statistics submitted to college data collection agencies. Further, the importance with which students regard certain college characteristics varies immensely from student to student. A wealthy student need not worry about cost, financial aid, or upward mobility, but these factors may be most important to a student who comes from a middle-class household. Instead of arbitrarily weighting the variables included in a rankings model and applying it to every student, a model more responsive to a student's individual preferences can help students make a more accurate decision when comparing multiple colleges. Lastly, emphasizing hierarchy in college admissions primarily benefits prospective students from privileged backgrounds—as described by McDonough et. al. in 1998:

> "The students who are using the rankings are precisely those students who have fine-tuned perceptions of what is important in choosing a college and who already know, and act on, notions of which institutions are best. Newsmagazine rankings are merely reinforcing and legitimizing these students' status obsessions... Newsmagazine college rankings are merely heightening a preexisting American obsession with reputations in a widespread, accessible way."

For these reasons, this paper proposes two ideas that attempt to overcome the issues caused by rankings models like those popularized by U.S. News. First, we build a model for the purpose of estimating the similarity between any two schools. The measure, henceforth referred to as a *similarity-score*, will be based on a number of institutional characteristics and be used to compare multiple colleges to some college of interest. The second idea is to make this model personalizable to an individual user's interests—an interactive model incorporated into a web-based application will allow users to prioritize different categories of college characteristics and adjust the weights of the similarity model accordingly. Both of these ideas directly address the problems perpuated by popular school-ranking models.

To address these goals, we constructed a dataset from two different data sources to effectively capture a wide spectrum of institutional characteristics ("variables") on a large number of schools. The principal data source was The Integrated Postsecondary Education Data System (IPEDS), a federally-sponsored data collection service to collect and distribute data on American schools receiving federal funding. Our second data source was Opportunity Insights' "Mobility Report Card," providing a more complete picture of postgraduate financial outcomes. These data were joined with the condensed IPEDS data to construct our final dataset. Our dataset ultimately contains 68 variables on 1,286 schools. See Section 2.1 for more information on our data acquisition procedure.

In order to perform quantitative analyses on this large number of school characteristics, we perform Principal Component Analysis (PCA) on subsets of the data embodying similar themes. PCA is a dimension reduction algorithm which transforms the dimensions of a space to new axes ordered by the variability of the data captured. This algorithm can be seen as a "squishing" process which approximates the distribution of schools based on their many characteristics into a condensed number of axes that contain most of the information captured by all of the variables. The

benefits of PCA are twofold: one use is for visualizations, where users can explore the approximate distribution of schools across 60+ variables for themselves. The second is distilling high-dimensional and varied data into a standardized "thematic" form, where differing numbers of original variables are distilled into single axes which approximate the high-dimensional distribution of schools.

We propose a similarity-quantifying algorithm tailored to prospective college students and institutional researchers interested in a subset of variables. This *preference-algorithm* uses PCA to reduce variables into themes (size, cost, location, mobility, diversity, exclusivity, and educational environment) that are single-dimension axes which approximate the more complex distribution. From here, a weighted euclidean distance is applied to the themes of interest between a chosen "school of interest" and all others, where schools with the smallest distance between them are deemed most similar. This feature allows users to specify both the institutional characteristics relevant to them, and to what extent they care about a certain theme via a slider which influences the weight each thematic characteristic is given.

We integrate this methods into a user-friendly web-based application built in *R Shiny*.[1] This tool serves to provide an interface for non-technical users to effectively use the available data to inform college decisions. Making use of the dataset resulting from Section 2.1, the application allows users to select from two to four school characteristics they find most important from a sidebar, as well as an optional set of schools of interest, and generates both an interactive plot and table. Both outputs use aesthetic elements to emphasize the schools of interest and situate them among similar schools. Schools displayed in the plot can also be hovered over to give a basic summary of that school's name and characteristics—this information will automatically be displayed if there are few enough points in the current pane. The accompanying table reacts to user input, both in the plot and sidebar, and displays the "raw values" of the characteristics of the chosen schools. The application also uses the total similarity-scores and preference similarity-scores to provide further information to the user. Altogether, the tool effectively synthesizes a data-centered and non-hierarchical approach for college exploration for prospective college students and institutional researchers.

[1]The app is freely available at `https://shiny.reed.edu/s/users/couchs/colleges/` with source code at `https://github.com/Reed-Statistics/College-Exploration`

Chapter 2

Methods

2.1 Data Acquisition

The data utilized in these analyses are provided by The Integrated Postsecondary Education Data System (IPEDS) and Opportunity Insights' "Mobility Report Card". IPEDS data is collected by the United States Department of Education's National Center for Education Statistics every two years and consists of fifty data sets across twelve survey categories for each year. Completion of the surveys is required for all institutions that participate in programs receiving financial assistance through Title IV of the Higher Education Act of 1965. For the methods outlined in this work, we use a subset of the variables collected in 2016.

Depending on the level of observation (some on schools, some from the student's perspective), each data set contains between 1,722 and 165,050 observations on seven to 647 variables. We found that this data source was the most comprehensive and complete of all publicly available data. IPEDS contains information for each school on basic characteristics, such as cost, admissions, and student demographics, and maintains relatively little missingness. To construct the IPEDS portion of our data set, we examined the codebooks of each IPEDS data set, recording variables with subjective relevance to the college research process. Clusters of college variables encoding similar information are henceforth referred to as themes, and encode information such as size, cost, location, diversity, mobility, exclusivity, and education style.

We queried the data using a codebase adapted from an existing R package (Bryer, 2018), and then collapsed the level of observation of each chosen variable such that each row represents a school. After binding the variables together into a single data set, we examined the levels of missingness in the collapsed variables. If a variable had an unsatisfactory level of completeness (>50%), we re-examined the IPEDS codebooks for variables that represented similar themes as the variables with high missingness, and repeated the process until all themes we desired to be represented in our data were reflected by multiple near-complete variables. After this process was complete, we ensured that schools in our sample reported a nonzero number of faculty and some form of admissions testing scores to filter out technical schools.

While the IPEDS data are highly comprehensive in documenting school charac-

teristics, information about post-graduate outcomes are less thorough. To account for this highly important element of a school's character more completely, we made use of Opportunity Insights' "Mobility Report Card." Opportunity Insights is a "non-partisan research and policy institute based at Harvard University focused on improving economic opportunity," and provides free data on, among other things, estimates of intergenerational mobility by school. We made use of the "Mobility Report Cards: Preferred Estimates of Access and Mobility Rates by College" data set in order to more comprehensively document post-graduate outcomes by school in our analyses. After joining these two data sources together (keeping all schools recorded in the IPEDS data, but not necessarily those in the Mobility Report Card), our finalized data set used for the following analyses contains 68 variables on 1,286 schools.[1]

2.2 Principal Component Analysis

Principal Component Analysis (PCA) is a common dimension reduction algorithm. For an extensive discussion on how PCA is designed, see *Introduction to Statistical Learning with Applications in R* (Hastie, Tibshirani, Witten, & James, 2013). In this project, PCA is employed to distill the information across several themes of variables into their own axes. This transformation is necessary to preserve meaningful euclidean distances, as variables such as undergraduate enrolled and total unrolled are very highly correlated. This would make school population count for more than a variable with singular encoding, such as number of professors. PCA solves this issue by collapsing highly correlated variables into single axes, which are called themes and used in a euclidean distance function turning preferences and theme encodings into similarity scores.

The PCA algorithm "re-phrases" the information of the observations into a transformed space where the axes encode the same information but in a different way. The new axes defining the space are sorted in terms of how much information they store. Using this technique, data in any number of dimensions can be approximated in $\mathbb{R}^2$ or $\mathbb{R}^3$. A common utility of PCA is to compress high-dimensional data into an approximation for visualization, such as 2 or 3 dimensions, which are the first axes of the transformed space.

[1] Our resulting data set and source code is freely available at PUBLIC project Git: (`https://github.com/Reed-Statistics/College-Exploration`)

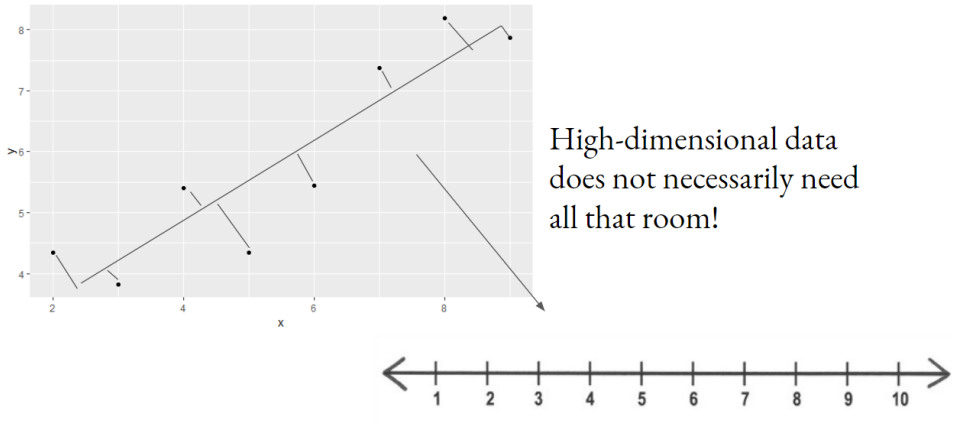


Figure 2.1: PCA projects data onto a hyperplane encoding the greatest variance to approximate the distribution of the data in one fewer dimension

The PCA algorithm works as follows:

1. Input a data set of n observations and p scalar variables
2. Normalize each variable to have a mean of 0 and a standard deviation of 1
3. Define a hyperplane q in $n-1$ dimensions along the axis of greatest variance
4. Project the observations onto q
5. Iterate p times

Principal Component Analysis iteratively projects the data along new axes in a reduced dimension which encode the highest amount of variance in the current space. The result of this is a number of variables ordered by the degree of variance of the axes on original data. This process squishes along correlated variables which can be approximated along one dimension.

Visualizations are a significant benefit of dimension reduction. Graphs with 2-4 unrotated variables leads to misleading inference due to potentially informative variables being left unseen. Visualizations using the first 2-4 principal components can be seen as an approximation of all variables of a high-dimensional data set, and can be highly explanatory if the components encode comprehensible themes. In college data, these most representative themes typically encode size, cost, test scores, and diversity, despite these themes not being expressed by a single column. Each theme is composed of collapsing one or more variables into a single axis. This makes themes all equally weighted for euclidean distance, but preserves approximative information on the distribution of colleges within that theme. Themes on the data are derived from collapsing correlated variables into an appoximative axis, shown in the axes below.

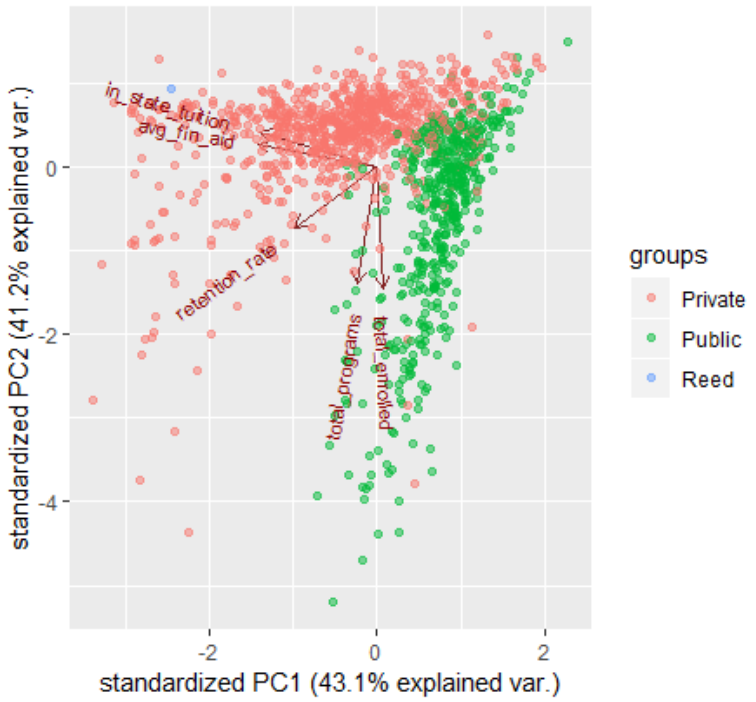


Figure 2.2: A demonstration of PCA, where red arrows indicate the contribution of the original variables into the PCA approximating axes. More parallel arrows indicate more correlated variables.

2.3 User-Preference Distance

User Preference Distance is geared towards a student or institutional researcher interested in researching academic institutions according to unique preferences. The user inputs a school of interest and scales the given themes according to their level of interest. The algorithm outputs a score for all schools based on the input levels of interest, and shows the highest scoring schools as well as further information that may interest them, where a higher score is a more similar school. This additional information is relevant data on the comparison schools such as ACT scores, cost, location, and institution type. The subsets created by PCA each encapsulate a "theme" of institutional characteristics, described in Figure 2.3:

Theme	Variables Included
Size	school size, dorm capacity, total enrolled, enrolled by gender, enrolled by race, first years enrolled
Cost	room and board, pct financial aid, avg financial aid, in state tuition, out state tuition
Exclusivity	ACT percentiles, percent admitted
Mobility	Mobility rates between bottom 20% family earners to top 20% and 1%
Diversity	proportion enrolled by gender, proportion enrolled by race
Education	highest degree, student faculty ratio, prop prof, prop instr, control, part time grad, part time ugrad, retention rate

Figure 2.3: Variable themes and their composition

Each of these themes is expressed as a 1-dimensional approximation of a number of variables from the original data set. This approximation is made by performing

PCA on just the subset of variables shown in Figure 2.3. In this way, redundant encoding is avoided even among uncorrelated variables in order to best preserve the interpretability of preference-scaling. Destroying redundancy through approximations is significant for the euclidean distance formula employed in the similarity algorithms which would be harmed by redundant variables. This is because similar types of variables would be double-counted in the score.

The algorithm works as follows:

1. Theme subsets are manually created which group multiple variables into subsets of size, cost, education, mobility, exclusivity and location.

2. PCA is performed on each subset separately.

3. The first principal component from each theme matrix is extracted and combined into a new matrix.

4. Modified euclidean distance is performed between the school of interest and all others. The modified Euclidean distance $\mathcal{D}$ uses scaling terms which indicate the level of user interest in each theme and returns a scalar

$$\mathcal{D}(\text{school}) = \sqrt{\alpha_1(x_1 - y_{i1})^2 + \cdots + \alpha_2(x_p - y_{ip})^2},$$

where $\alpha_i \in (0, 1)$ represents the degree of user interest in the ith theme. The terms x_i and y_{ip} represents the pth variable of the reference school and comparison school.

5. The similarity score is scaled to range from 0 to 100 and reported to the user along with additional information on the comparison schools.

The scores are returned on $[0, \infty)$, where 0 is the school itself (zero distance, perfectly similar). These are transformed to a more interpretable percentage scale, where the school of interest is 100% similar to itself, and a somewhat similar school is 80% similar. For example,

$$\mathcal{S}_i = 100 * (1 - \frac{\hat{x}}{\max \hat{x}}),$$

where $\hat{x}$ is the list of all scores and $\mathcal{S}_i$ is the ith schools percentage of similarity to the reference school.

Chapter 3

The College Exploration App[1]

In order to make our similarity-scoring algorithms approachable for all users, we constructed a web application in *R Shiny* so that the data and methods resulting from this study are more easily accessible. Making use of the data detailed in Section 2.1, the application centers user preferences and interests in an intuitive interface.

3.1 Basic Plotting

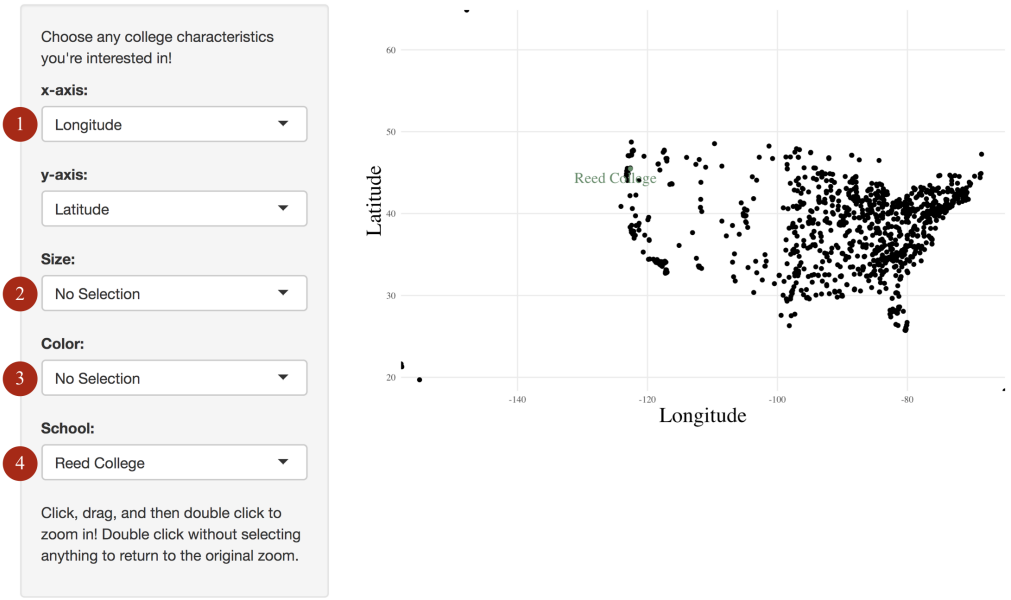


Figure 3.1: The default home screen of the app is the basic plotting portion of the app, where users can visualize 2 to 4 school characteristics.

Upon starting the application, the user first sees the plotting tool. The user can choose to visualize between two and four school characteristics in addition to the option of

[1]The app is freely available at `https://shiny.reed.edu/s/users/couchs/colleges/` with source code at `https://github.com/Reed-Statistics/College-Exploration`

choosing a number of schools of interest to accentuate in the visualizations.

1. **Axes**: The user can select any of the numeric variables in the data set to be the x and y axes of a scatterplot. The axis limits automatically readjust upon selection of new variables from the dropdown menus. The user can zoom into any region by dragging and double clicking a region. If the plot is zoomed in such that there are twenty or fewer schools, each school will be labeled with its name. To reset the plot zoom, the user can double click anywhere in the pane.
2. **Size**: The size of the plotted points can be mapped to any numeric or ordered factor variable in the dataset. Upon selection of a variable, the subset of observations currently shown in the window are ranked on the variable of interest. Though this decision means that the true difference in parameter value between observations is not possible to deduce from the plot, this transformation is still useful. Since many variables in the dataset are heavily skewed, mapping their raw values to size often results in the range of sizes plotted being unhelpful in showing differences in characteristics between schools for all but extreme observations (on the outer tails of the distribution of the variable). By transforming these values to take a uniform distribution, the *order* among the schools is more clearly apparent in the visualizations. Further, if a user desires to see the raw values of the chosen variable, the interactive tables detailed in Section 3.2 make these values readily accessible to mitigate the downside of the visual transformation.
3. **Color**: Any variable in the dataset can to be mapped to the color variable. In determining the color palette to map the values to, the application makes use of the column type of the selected variable. Numeric variables are assigned continuous diverging palettes with the reference school placed at the center of the spectrum, ordered factor variables are assigned ordinal categorical palettes, and factor/character variables are assigned categorical palettes.
4. **Schools of Interest**: The user also has the option to choose a set of schools of interest to highlight in the visualizations. If no selection is made for the color variable, the schools are labeled in a different color than other schools in the pane (in the case that any other schools are labeled). When a selection for the color variable is made, the newly chosen color mapping overrides the color denoting whether the school is chosen or not, but the school is labeled in a different color than the other labeled schools. Similarly as with color, the size of the point is made larger than other points in the pane unless the size variable is chosen, in which case color is utilized to accentuate the point.

3.2 Interactive Table

Show 10 entries Search:

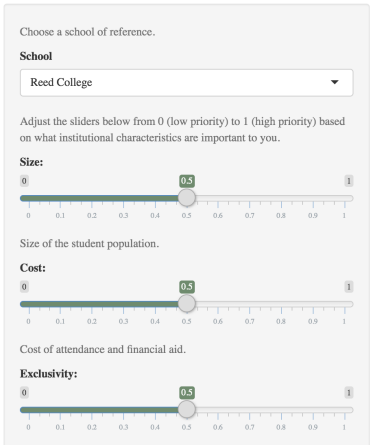
School Name	State	School Type	Carnegie Classification	Longitude	Latitude
All	All	All	All	All	All
Lewis & Clark College	OR	Private not-for-profit	Liberal arts	-122.672478	45.450547
Reed College	OR	Private not-for-profit	Liberal arts	-122.63064	45.480376
Concordia University-Portland	OR	Private not-for-profit	Universities	-122.637944	45.568242
Sonoma State University	CA	Public	Universities	-122.673261	38.341023
Portland State University	OR	Public	Universities	-122.683553	45.511229
Southern Oregon University	OR	Public	Universities	-122.694034	42.186467
Dominican University of California	CA	Private not-for-profit	Universities	-122.514654	37.980014
Saint Martin's University	WA	Private not-for-profit	Universities	-122.815441	47.04075
Western Washington University	WA	Public	Universities	-122.484873	48.737236
San Francisco State University	CA	Public	Universities	-122.477905	37.721345

Showing 1 to 10 of 1,286 entries Previous 1 2 3 4 5 … 129 Next

Figure 3.2: The interactive table portion of the app is integrated with the basic plotting portion, giving raw data values of the plotted points.

Though the minimalistic design of the plots described in Section **??**sec:basic-plotting) allows for expressive and intuitive interfacing with the data, it restricts the users ability to effectively interpret raw data by the plots alone. In order to supplement the plots with accurate raw data, we implemented a reactive table that is displayed below the plot. By default, the table displays only the variables selected for plotting with the chosen schools placed at the top of the table. After the chosen schools, remaining schools are arranged by their distance along the x and y axes to the mean x and y values of the chosen schools so that information on schools similar to those chosen is made more prominent. However, the user can sort the rows in the table by any of the plotted variables in addition to searching for a school using a search box in the top right. This feature offers users a data-centered approach to college exploration synthesizing visual intuition and quantitative transparency.

3.3 Similarity Scoring



Show 10 entries Search:

Name	Similarity Score	State	Type	Carnegie Classification	Most Similar Trait	Second Most Similar Trait
Reed College	100	OR	Private not-for-profit	Liberal arts		
Whitman College	91.63	WA	Private not-for-profit	Liberal arts	Size	Mobility
Pomona College	87.83	CA	Private not-for-profit	Liberal arts	Size	Cost
Macalester College	86.84	MN	Private not-for-profit	Liberal arts	Exclusivity	Education
Westmont College	86.31	CA	Private not-for-profit	Liberal arts	Size	Mobility
Soka University of America	86.23	CA	Private not-for-profit	Specialized	Mobility	Size
St Olaf College	84.15	MN	Private not-for-profit	Liberal arts	Education	Diversity
Lawrence University	83.51	WI	Private not-for-profit	Liberal arts	Size	Mobility

Figure 3.3: The similarity scoring portion of the app utilizes the algorithm described in Section 2 to compare a chosen school with all others in our data.

In addition to providing basic plotting tools and interactive tables, the web application also includes methods implementing the similarity-scoring algorithms described in Sections 2.2 and 2.3. The application provides a set of sliders allowing the user to indicate the relative emphasis they might place on different school characteristics, runs the user-preference algorithm described in Section 2, and returns a table displaying basic characteristics (as well as similarity scores to the chosen school) of the most similar schools. Further, for each similar school, the application presents the two most important themes in determining similarity to the chosen school so that users can understand the factors that most influenced the resulting score.

3.4 Network Visualization

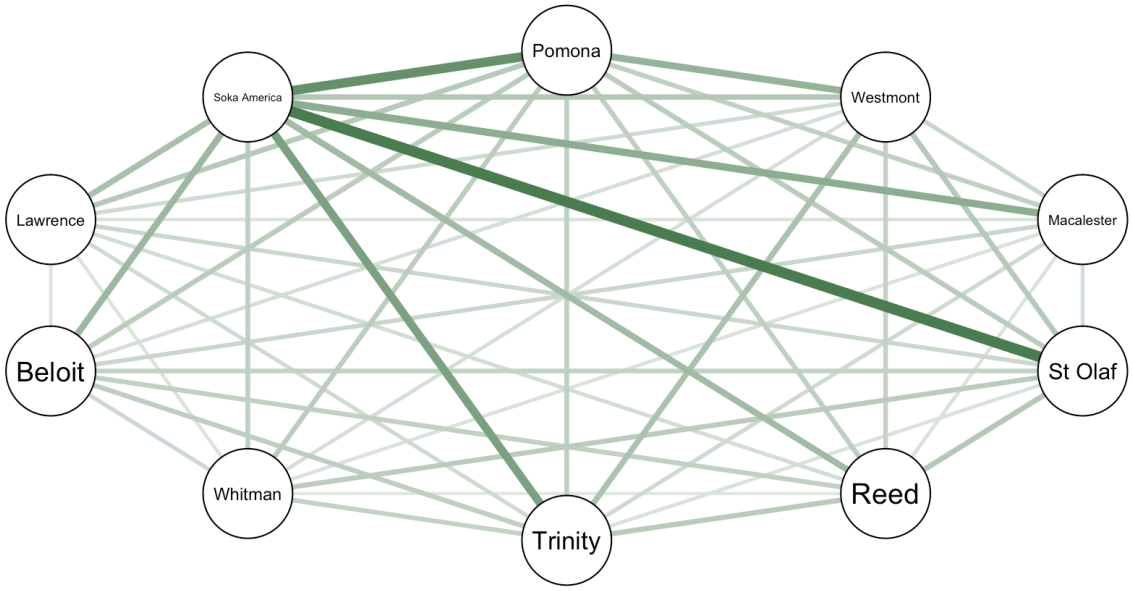


Figure 3.4: The algorithm described in Section 2 is also utilized to generate a network visualization depicting pairwise similarity scores for the top 10 schools most similar to the reference school.

Along with the table providing similarity scores, a network visualization constructed based on pairwise similarities of the top 10 schools in the table is displayed. Higher similarity scores are represented by thicker and more opaque lines, while lower similarity scores are depicted with thinner and more transparent connecting edges. Just like in the basic plotting setting, the network visualization aims to provide intuition about school similarity rather than to provide exact values. This network visualization gives meaningful insights into the distribution of schools according to user preference at the cost of being able to approximate raw data directly from the plot. However, as in the basic plotting setting, making use of the data table below the interface provides the user with the true similarity scores to the reference school.

Chapter 4

Discussion

Our models and tools heavily depend on the quality of data provided by IPEDS and Opportunity Insights. At times, the quality of the data is suspect. For example, some schools listed that they accepted a larger number of students than the number of students that applied for enrollment. In addition, some schools had unexpectedly low acceptance rates, and our model thus asserted that these schools were similar to institutions generally regarded as elite. The more egregious of these data quality issues, as mentioned above, are less problematic than the smaller and harder-to-detect errors in this data, as users will easily see past the former.

Another concern about data quality is the incentive for school officials responding to the IPEDS surveys to massage their numbers. Although the goal of this project is to propose an alternative to rank-based college comparisons, no matter how we approach the data, we cannot eliminate the problem that schools are consciously aware that their responses will have a direct impact on the way their school is evaluated by prospective college students. Schools can misreport their statistics by rounding up, choosing to interpret a survey question in the way that portrays their school in the most positive light, or reporting that statistics that would be disadvantageous to the school are missing.

The large dimensionality of the data meant that, for one, it was impossible to visualize all variables at once, but also inadvisable to provide every available form of information to the user. Finding a balance between building a flexible app that encoded a lot of the information that might be important to the user, and building a more stable one that was less overwhelming but omitted some useful features was a significant challenge. The two tools seek to remedy this discrepancy by providing both broader summaries or features and high-specificity exploration through tables and visuals. Again, these tools also introduce a probblem balancing between making the app capable of making complex visualizations while still being easy to use. Much effort has been made making the app both approachable to a new user, and powerful enough to be useful to a savvy user. Ultimately, these concerns of project domain are alleviated though thorough documentation and guides. To maintain a usable yet powerful tool, the balance between functionalities and usability must be navigated through ample instruction and approachability. Pop-ups, informative panes, clear labels, and accessible documentation make the difference between a cryptic tool and a

powerful one.

There are many potential directions for further work on this tool and similar ones. Rather than geographical distance between schools, it would be helpful if instead users could describe what areas of the country they want to focus on. One way to do this would be to limit schools that are within a certain radius of a location the user provides. This way, being closer does not make the school more similar, but rather they limit the number of schools and treats distance as equal for all of them. Another option is to code for variables that could be more useful than the ones we used. A net cost variable that factors in tuition the student should expect to pay (whether they are an in-state or out-of-state student), financial aid, and cost of living would be an excellent way to provide useful information in one column of data that, in our case, uses multiple columns. Further, determining a way to identify conventionally recognized peer schools, as opposed to schools that our tool finds similar based on quantitative measures, would be useful for those using the app from the perspective of an institutional researcher at a single college. Each of these proposed features would extend the users ability to compare schools of interest without emphasizing hierarchy.

Nevertheless, the goal of this study was to provide an alternative to rank-based college evaluation, and we have succeeded in creating an application that comprehensively compares schools of interest using a data-centered approach.

References

Bryer, J. (2018, December). R package for Interfacing with the Integrated Postsecondary Education System (IPEDS). Retrieved from https://github.com/jbryer/ipeds

College Admissions Scandal. (2019). *The New York Times*. Retrieved from https://nyti.ms/2UDavdU

Furukawa, D. T. (2011). College choice influences among high-achieving students: An exploratory case study of college freshmen.

Hartocollis, A. (2019). College admissions: Vulnerable, exploitable, and to many americans, broken. *The New York Times*. Retrieved from https://www.nytimes.com/2019/03/15/us/college-admissions-problems.html

Hastie, T., Tibshirani, R., Witten, D., & James, G. (2013). An introduction to statistical learning with applications in r. New York: Springer.

How U.S. News Calculated the 2019 Best Colleges Rankings. (2019). *US News & World Report*. Retrieved from https://www.usnews.com/education/best-colleges/articles/how-us-news-calculated-the-rankings

Kim, J. (2018). The Functions and Dysfunctions of College Rankings: An Analysis of Institutional Expenditure. *Research in Higher Education*, *59*(1), 54–87. http://doi.org/10.1007/s11162-017-9455-1

Luca, M., & Smith, J. (2013). Salience in Quality Disclosure: Evidence from the U.S. News College Rankings. *Journal of Economics & Management Strategy*, *22*(1), 58–77. http://doi.org/10.1111/jems.12003

Lydgate, C. (2018). Reed and the Rankings Game. *Reed Magazine*.

McDonough, P. M., Lising, A., Walpole, A. M., & Perez, L. X. (1998). COLLEGE RANKINGS: Democratized College Knowledge for Whom? *Research in Higher Education*, *39*(5), 513–537. http://doi.org/10.1023/A:1018797521946

Medina, J., Benner, K., & Taylor, K. (2019). Actresses, business leaders and other wealthy parents charged in u.s. College entry fraud. *The New York Times*. Retrieved from https://nyti.ms/2F8Km15

Ortagus, J. C. (2016). Pursuing Prestige in Higher Education: Stratification, Status,

and the Influence of College Rankings. *College and University; Washington*, *91*(2), 10–19. Retrieved from https://search.proquest.com/docview/1806232416/abstract/6485E7832DA5458DPQ/1

Stecklow, S. (1995). Colleges inflate sats and graduation rates in popular guidebooks. *Wall Street Journal*, *5*.

Stephens, B., & Collins, G. (2019). The kids aren't all right. *The New York Times*. Retrieved from https://www.nytimes.com/2019/03/19/opinion/college-scandal-massacre-guns.html