Ziyang Shao

Honor Thesis

Advisor: Professor Ji Zhu

24 April 2020

College Ranking Based on Pairwise Preferences

Abstract

This paper adopts the Bradley-Terry model and Newman's community detection algorithm to infer students' choice preferences on universities in the United States as an indication of school reputation and to determine influential factors for their decision-making. The framework of ranking is based on college cross-admit comparison data from Parchment, revealing the percentage of students choosing one school over another while receiving offers from both. Community detection is applied to identify different school groups in applications. We found that for high achieving students, school reputation outweighs geographical disturbance, while typical students prefer not to travel too far for college. We also notice that colleges in California and New York are generally considered together with nationwide colleges rather than in a regional, local college network.

Keywords: school-ranking, Bradley-Terry model, network analysis, community detection, student preference

Introduction

Going to college is a life-changing decision for most students, and the school attended can drive one's life in very different directions. School ranking is a helpful tool for students to decide where to apply and to make final decisions. Popular ranking methodologies today are concerned with a wide variety of factors, and assign different weights to different factors to obtain a final evaluation. While their methods to calculate scores for each factor can differ, the main factors concerned always include retention and graduation rate, social mobility, faculty performance, financial resources, and employer reputation. However, schools' reputations among students are always neglected, though it is an important metric beneficial in learning how other students choose one school over another, especially during the final decision period.

In this context, investigation into the school application pattern of past students could provide useful information for future high school students. By constructing a Bradley-Terry preference ranking out of the winning rate of one school over another, and at the same time detecting clusters of schools based on students' application behavior, we obtain several ranked groups of schools. Several application behavior patterns are also identified for future students to adjust and reflect on their application process.

Data

The preference data is collected from Parchment, a widely adopted digital credential service platform mainly used to exchange transcripts between students and schools. The website claims to have "exchanged more than 30 million transcripts and other credentials globally" for "millions of people and thousands of schools and universities," and have collected a database of over 2,044,079 acceptances at hundreds of colleges in the US.

The data we use is from their well-known “Side by Side College Comparison”, where the user can choose two schools and see a percentage for each school as the revealed preference. For each school’s percentage, the denominator includes all members who were admitted to both of these schools. The numerator includes those students who chose a given school. A confidence interval at the 95% level is also represented, calculated by Wilson’s method:

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Unreliable statistics, resulting from insufficient data size, are indicated. We crawled and curated valid pairwise comparisons for 839 schools, altogether 41296 pairs, including school names, winning rates and confidences. We then reversed Wilson's formula to estimate the total matchups between the two schools (number of students who were admitted to both of these schools and attended one of them) and the exact number of students choosing one given school.

The following figures give an overview of the dataset, namely “estimated matchups”. Figure 1 shows the log of the sum of matchups for each state, indicating the number of offers recorded in each state when it is not the only one for a student. Figure 2 exhibits the average number of matchups for each school in each state, excluding in-state matchups (receiving offers from the same state) and out-state matchups respectively. As exhibited, among all states, California and Michigan have the highest number of matchups. Also, it is easy to see that the average in-state matchup is much larger than out-state numbers, indicating that universities tend to make offers to students in the same state.

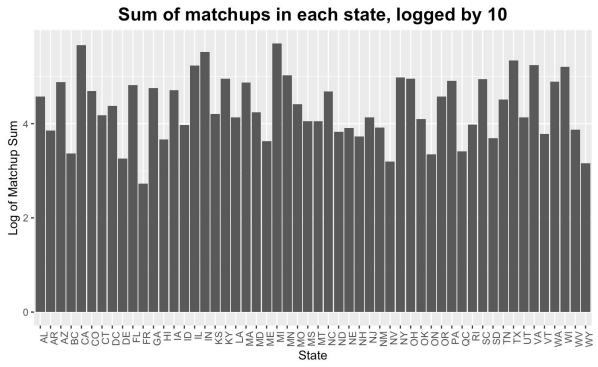

Figure 1, Total Matchups in Each State

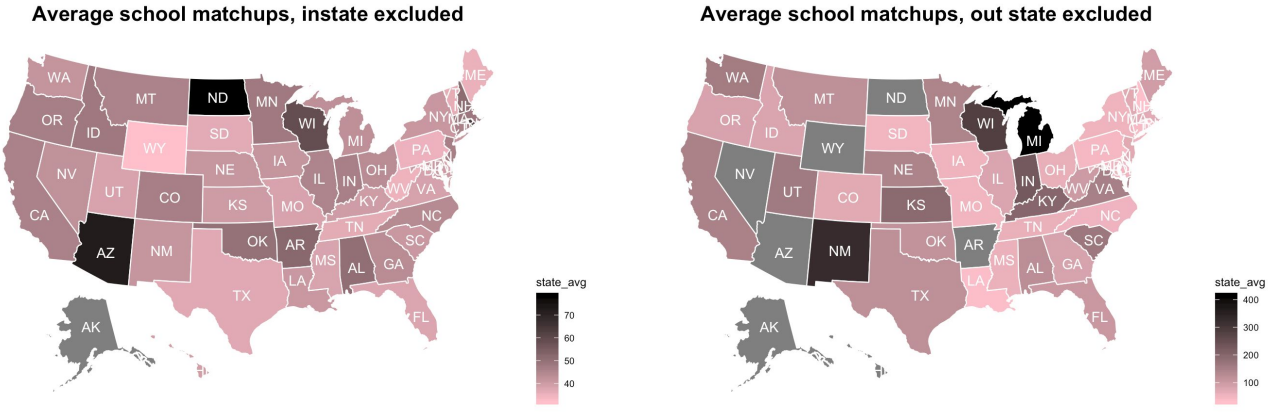

Figure 2, School Average Matchups in Each State

Next, we created an undirected network from the dataset for network analysis. In this network, each vertex represents a school, and if between two schools there exists matchups, an edge is created between the two vertices. Figure 3 illustrates the number of schools with which a given school has at least one matchup, counted as degree. It is revealed that most schools have connections with around 30 to 70 schools.

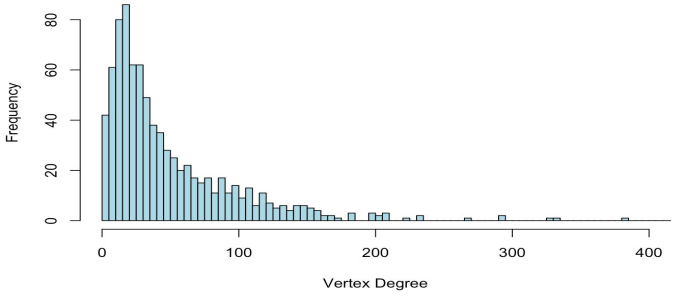

Figure 3, Histogram of Number of Schools with Which One Has Matchup

Among the schools, Texas A & M University has the largest matchup number, 794. We found that it is applied together with almost every other college for at least 20 students each. Following Texas A & M University are University of Washington, Columbia College-Chicago, University of Michigan, and Purdue University, with matchup numbers being 385, 332, 327, and 294 respectively. Except for Columbia College-Chicago, the other four colleges with leading matchup numbers are well-known flagship public universities, which are probably common choices in one's application and also make a considerable number of offers. Note that Columbia College-Chicago, on the contrary, is a low-ranking private institution with a high tuition fee of $157,446. Its large matchup number probably results from a profuse willingness to set a low entrance bar and make more offers. According to Niche, its acceptance rate is 87%, significantly higher than 23% (University of Michigan), 49% (University of Washington), 58% (Purdue University-Main Campus), and 68% (Texas A & M University).

Methods and Findings

A. Application of the Bradley-Terry model to obtain college preferences

To understand students' preferences when receiving multiple offers, we consider the Bradley-Terry model (Bradley & Terry, 1952). This model of paired comparison has been widely and effectively used in ranking stimuli from paired comparisons since proposed, especially in situations where it is difficult to quantify differences among the items. "Desirable properties of paired comparisons, in comparison with other rating methods, include the minimal constraints placed on the response behavior of individuals and the wealth of information that can be obtained regarding individual preferences as well as regarding the

perceived similarity relationships between choice stimuli" (Satoshi Usami, 2010). In the

$$P_{ij} = \frac{\theta_i}{\theta_i + \theta_j},$$

Bradley-Terry model, the probability of choosing a stimulus *i* over *j* is expressed as: where θ*i* is a positive-valued parameter which might be viewed as a representation of stimulus *i*'s ability. It can also be expressed as: logit [P (*i* beats *j*)] = λ*i* − λ*j,* where λ*i* = log(θ*i*).

Many extensions of the Bradley-Terry model have been developed. For example, Davidson (1970) proposed a solution to situations where no preference is allowed; Causeur and Husson (2004) introduced a 2-dimensional extension to eliminate the constraint of linear scale of merit and accommodated situations where merits are not transitively related. Firth (2005) implemented the classical Bradley-Terry model in R and published his R package *BradleyTerry*, which is adopted in this research. Figure 4 demonstrates the distribution of school preference scores estimated by the Bradley-Terry model. Table 1 compares the top 25 schools produced by our model and the corresponding USNews ranking.

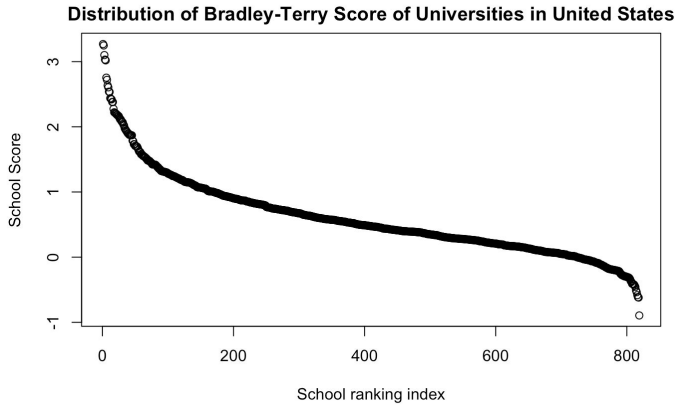

Figure 4, Distribution of Bradley-Terry Score of U.S. Colleges

Bradley–Terry Ranking	University	Bradley Terry Score	US News Ranking
1	Massachusetts Institute of Technology	3.269608	3
2	Stanford University	3.248204	6
3	Harvard University	3.100589	2
4	Princeton University	3.031410	1
5	Yale University	3.019759	3
6	Caltech	2.754171	12
7	University of Pennsylvania	2.720556	6
8	Brown University	2.635197	14
9	Duke University	2.605886	10
10	University of California, Berkeley	2.538666	22
11	University of Michigan – Ann Arbor	2.534638	25
12	University of Chicago	2.438115	6
13	Columbia University in the City of New York	2.427906	3
14	Pomona College	2.421107	NA
15	University of Notre Dame	2.380987	15
16	University of California, Los Angeles	2.380363	20
17	Dartmouth College	2.275693	12
18	United States Air Force Academy	2.222394	NA
19	University of Virginia	2.217045	28
20	University of Southern California	2.205737	22
21	Northwestern University	2.199596	9
22	Swarthmore College	2.184780	NA
23	United States Military Academy	2.181362	NA
24	Washington University in St. Louis	2.167098	19
25	Rhode Island School of Design	2.163719	NA

Table 1, Comparison of Bradley-Terry Ranking and U.S. News Ranking for Top Colleges

As can be seen from the results by our paired-comparison preference model, research-oriented universities, such as Massachusetts Institute of Technology, Stanford University, California Institute of Technology, University of California, Berkeley, University of Michigan gain more preferences than indicated in the USNews ranking. Two military schools, United States Air Force Academy and United States Military Academy, are also revealed to be competitive. Columbia University, a well recognized top university in most ranking systems, surprisingly, came out as not preferred in comparison to other first-tier schools, dropping by 10 places compared with USNews ranking. It is also worth noting that two liberal arts colleges, Pomona College and Swarthmore College, rank high by our model.

As previously noted, research-oriented universities are favored compared with their USNews ranking, generally gaining more preferences from students. To explain such advantages, we consider factors influencing the college decision process, on the basis that the "core of college choosing is to attend a high quality college or university" (George, Suzanne, and Charles, 2001). Recently, Connie and Rahman (2019) identified the program, university reputation, employment opportunity, pricing, security, education and campus facilities, and location and peers as main factors affecting students' choices.

B-T rank	USNews rank	Difference	School	Best Known For (according to Google)
1	3	+2	MIT	Engineering; Physical Sciences
2	6	+4	Stanford	Computer and Information Sciences and Support Services
3	2	-1	Harvard	Social Sciences; Mathematics; History
4	1	-3	Princeton	Social Sciences; Engineering
5	3	-2	Yale	Political Science; History
6	12	+6	Caltech	Engineering; Physical Sciences
7	6	-1	UPenn	Business; Management; Marketing
8	14	+6	Brown	Computer Science; Econometrics and Quantitative Economics
9	10	+1	Duke	Computer Science; Econometrics and Quantitative Economics
10	22	+12	UCB	Engineering; Biological and Biomedical Sciences
11	25	+14	UMich	Computer Science; Chemistry
12	6	-6	UChicago	Social Sciences; Mathematics and Statistics
13	3	-10	Columbia	Social Sciences; Engineering

Table 2, Ranking Comparison of Top Schools and Their Best-Known Majors

BROAD CATEGORY	AVERAGE SALARY
Computer and information sciences and support services	$90,350
Engineering	$82,242
Engineering technologies	$81,041
Transportation and materials moving	$78,467
Business, management, marketing, and related support services	$71,904

Source: Summer 2019 *Salary Survey*, National Association of Colleges and Employers. Note: Only disciplines with 50 or more salaries reported are included.

Table 3, Starting Salaries by Discipline for Class of 2018 Graduates

(Data Source: Summer 2019 Salary Survey)

Table 2 indicates that the favored schools tend to share a common attribute in their most well-known majors: they all exhibit an advantage in some tech-based area such as engineering, physical science and computer science. As such majors promise more job opportunities and a better future income level (Table 3), we conjecture that the observed preferences in Bradley-Terry ranking are job-oriented.

As for the United States Military Academy and United States Air Force Academy, we found that except for rivals against Yale and Princeton, students choose one of the two military academies over other colleges in 77% of comparisons. Such students tend to have a specific proclivity to a military type school and probably have prepared for special requirements of such schools, so unless there is a competing offer from a world-prestigious university, the military academy stays as their first choice.

B. Application of Newman's community detection algorithm to identify school clusters

To touch on comparable schools, we perform network analysis on matchups between schools to extract which ones are usually considered together during the application process, thus shedding light on student's selection in alternative schools during application. Specifically, we apply the modularity algorithm (Newman, 2006) to identify community structures in the network using eigenvectors of a so-called modularity matrix, which is created from pair-comparisons between universities. This method detects modules in networks, defined as "groups of vertices with a higher-than-average density of edges connecting them." Newman (2006) approaches this problem by maximizing a benefit function over possible divisions of the network and defines the benefit function Q, named "modularity", to be:

Q = (number of edges within communities) − (expected number of such edges)

This maximization problem can be written in terms of the eigen-spectrum of a matrix, and Newman (2006) proposed three matrix-based algorithms. The first is a method utilizing the leading eigenvector, which can only divide the network into two modules; the second is a generalization of the leading eigenvector method to extract information from eigenvectors other than the leading one of the modularity matrix; the last one extends the second method into a vector partitioning algorithm to accommodate negative eigenvalues. Subsequently, Newman (2006) proposed a repeated subdivision approach for detecting more than two communities to better cater to real-world networks, which often contain multi communities. However, Newman (2006) also mentioned that while this iterative method appears to work well in practice, a more satisfying approach would be to work directly from the modularity of the complete network. He commented on the standard technique of k-means clustering based on group centroids applied by White and Smyth (2005) and pointed out in future development, it might be a choice "if applied to the centroids of the end-points of the vertex vectors."

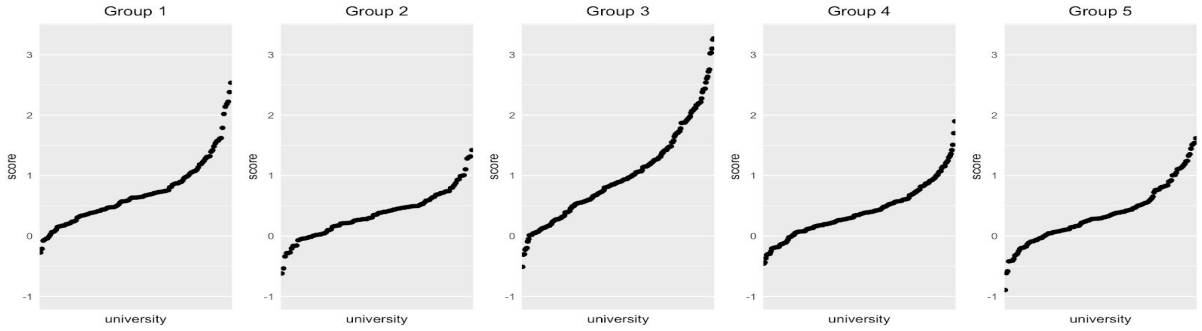

Figure 5, Distribution of B-T Score of U.S. Colleges After Community Detection

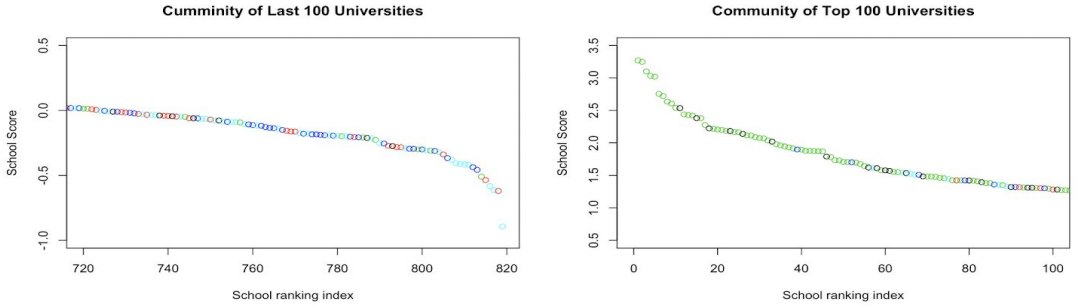

Figure 6, Last and First 100 Universities in B-T Ranking and Their Community Belongings

Note in Figures 5 and 6, top schools in Bradley-Terry (e.g. schools denoted by green circles) tend to cluster better than bottom-ranked schools, but overall, we do not observe that schools of similar rankings tend to cluster together.

university	score	university	score	university	score
University of Michigan – Ann Arbor	2.534638	Truman State University	1.4218021	Massachusetts Institute of Technology	3.269608
University of Notre Dame	2.380987	Kansas City Art Institute	1.3161974	Stanford University	3.248204
United States Air Force Academy	2.222394	Fort Hays State University	1.3024239	Harvard University	3.100589
United States Military Academy	2.181362	University of Washington–Tacoma Campus	1.2812702	Princeton University	3.031410
United States Naval Academy	2.138149	Illinois Wesleyan University	1.1030209	Yale University	3.019759
University of Texas at Tyler	2.019030	Cornish College of the Arts	1.0055193	Caltech	2.754171
Georgia Institute of Technology–Main Campus	1.788983	Augustana University	0.9978262	University of Pennsylvania	2.720556
Virginia Tech	1.620779	University of Redlands	0.9914141	Brown University	2.635197
Washington and Lee University	1.609715	Chapman University	0.9389906	Duke University	2.605886
United States Coast Guard Academy	1.577546	University of Missouri – Columbia	0.9269599	University of California, Berkeley	2.538666
University of Florida	1.568971	Edgewood College	0.8758516	University of Chicago	2.438115
The University of Texas at Brownsville	1.534443	Johnson County Community College	0.8379856	Columbia University in the City of New York	2.427906
University of Texas at San Antonio	1.483097	Berklee College of Music	0.8017161	Pomona College	2.421107
Virginia Military Institute	1.419938	Rider University	0.7939120	University of California, Los Angeles	2.380363
Ohio State University–Main Campus	1.393974	University of Central Missouri	0.7428355	Dartmouth College	2.275693
University of Delaware	1.320215	Grand Canyon University	0.7402257	University of Virginia	2.217045
University of Texas at Arlington	1.309438	Lawrence University	0.7392028	University of Southern California	2.205737
Clemson University	1.308301	Southern Utah University	0.7233278	Northwestern University	2.199596
Wake Forest University	1.280545	North Greenville University	0.7130032	Swarthmore College	2.184780
Texas A & M University–College Station	1.244103	Loyola University – Chicago	0.7116933	Washington University in St. Louis	2.167098
Florida Agricultural and Mechanical University	1.219089	Brigham Young University–Idaho	0.6936668	Rhode Island School of Design	2.163719
University of Texas at Dallas	1.189709	Lake Forest College	0.6881326	Bowdoin College	2.114676
United States Merchant Marine Academy	1.184472	Westmont College	0.6818857	Williams College	2.113282
University of South Carolina–Columbia	1.135940	Clark Atlanta University	0.6477113	Harvey Mudd College	2.088508
Saint Mary's College	1.098591	North Park University	0.6417168	Amherst College	2.073081
Florida State University	1.077100	Missouri State University	0.6223108	Cornell University	2.069044
Augusta State University	1.071630	Marymount Manhattan College	0.5959742	Rice University	2.037357
University of Alabama at Huntsville	1.059467	Weber State University	0.5891025	Claremont McKenna College	1.978942
Concordia University – Texas	1.052408	Beloit College	0.5854972	Vanderbilt University	1.961239
Rhodes College	1.028732	University of Iowa	0.5559670	Wellesley College	1.945121

university	score	university	score
University of Wisconsin – Madison	1.8979224	Parsons School of Design	1.6168820
Baker College of Flint	1.7009491	Midway College	1.5421537
Purdue University–Main Campus	1.5081496	University of the Cumberlands	1.5267227
Kendall College	1.4213609	Berea College	1.5096355
Michigan State University	1.3581516	Pennsylvania State University–Penn State Altoona	1.4448746
Hillsdale College	1.3201672	Pennsylvania State University–Penn State Harrisburg	1.3527720
University of Wisconsin – Superior	1.3000854	Christopher Newport University	1.3274754
Illinois State University	1.2447315	University of Maryland Eastern Shore	1.2432076
Saint Johns University	1.2228655	Fashion Institute of Technology	1.2359802
University of Illinois at Springfield	1.2067036	University of North Alabama	1.1961799
Indiana University – Northwest	1.1543593	Clayton State University	1.1721506
Purdue University–Calumet Campus	1.1524676	Kentucky State University	1.1484015
College for Creative Studies	1.1459858	Virginia Commonwealth University	1.1299028
University of Illinois at Chicago	1.1254492	Kentucky Wesleyan College	1.1180434
Illinois Institute of Technology	1.0723621	Morehouse College	1.1094057
Indiana University – Bloomington	1.0666059	Jacksonville State University	1.1050956
University of Wisconsin – Milwaukee	1.0532218	West Chester University of Pennsylvania	1.0478276
Evangel University	1.0116435	University of Pittsburgh at Johnstown	1.0093644
Ave Maria University	1.0066466	Kalamazoo College	1.0093025
Baker College of Owosso	0.9856520	Southern University and A & M College	1.0029969
University of Northwestern Ohio	0.9590117	Howard University	0.9221889
Elmhurst College	0.9487340	University of Toledo	0.9045319
Aurora University	0.9333975	Spelman College	0.9003055
DePaul University	0.9241591	Alderson–Broaddus College	0.8395190
University of Wisconsin – Stout	0.9162699	University of South Carolina–Aiken	0.8332258
Northwood University–Michigan	0.9098283	West Virginia University Institute of Technology	0.8247372
Indiana University – Southeast	0.8823019	Lindsey Wilson College	0.8200055
University of Wisconsin – Platteville	0.8710491	Wilmington College	0.8101426
Baker College of Muskegon	0.8695403	University of South Carolina–Upstate	0.8094386
Columbia College–Chicago	0.8675852	Tennessee State University	0.7934135

Table 4, Top Colleges in Each Community

In the detected communities, it is observed that the best universities and best liberal arts colleges are clustered in one module (Group 3). This phenomenon agrees with the claim made by Bradshaw, Espinoza and Hausman (2001), that academically talented students have a tendency to take reputation as first concern. They describe the students interviewed to "enter the college selection process with predispositions as to the kind of colleges they would consider attending, not whether or not they would attend college, and these predispositions shape their later activities." As they pointed out, "these predispositions include a desire to attend a prestigious college, a desire to enroll in a highly ranked academic program, and the expectation that they would receive significant scholarships." Moreover, the first tier liberal arts colleges being in this cluster reveals that for high achieving students, college type is not outweighing school rankings, and they are willing to be flexible for the type of education received to accommodate their predispositions.

Another identified community (Group 1) includes schools such as University of Michigan, University of Notre Dame, University of Texas and Georgia Tech, mostly consisting of well-recognized public universities in the nation. A third group (Group 5) contains colleges that are either Christian or liberal arts colleges with an emphasis on the art industry.

As can be seen in Figure 7, the location of school seems to affect students' selections during the college application process. As denoted, members of the third module appear to be distributed nationwide; in contrast, the second community, whose Bradley-Terry scores are generally the lowest among all schools, also show a tendency of dispersing across the nation. In both graphs, California is the state with the largest number of schools in the community of application list, followed by Pennsylvania and New York. Compared with the overall college distribution, we can see that the three states have more colleges. The abundance of in-state

choices are not restricting the students to stay at their residence but driving them to explore more on nationwide options, and we hypothesize that students in the three states tend to be more of "anywhere people" rather than "somewhere people." On the other hand, it is also possible that the copiousness of educational resources in these states attract out-of-state students, and combined with ample job opportunities, the attraction level outweighs geographical concerns and tuition fee discount of staying at home. Overall, colleges in these states tend to be broadly considered with across-country universities instead of inside a regional network. In groups 1, 4, and 5, there are clearly centers of students' choices: Texas, the Great Lakes region, and the South Atlantic region, respectively. In contrast to the first community, groups 4 and 5 are more concentrated around their respective centers. For example, group 4 centers around Michigan, Illinois, Wisconsin, Indiana, where there are large public universities (such as the Michigan State U, University of Illinois Chicago, U Wisconsin, and Purdue). . The concentration indicates that many students in these areas tend to consider a tighter range of colleges by putting more emphasis on their locations.

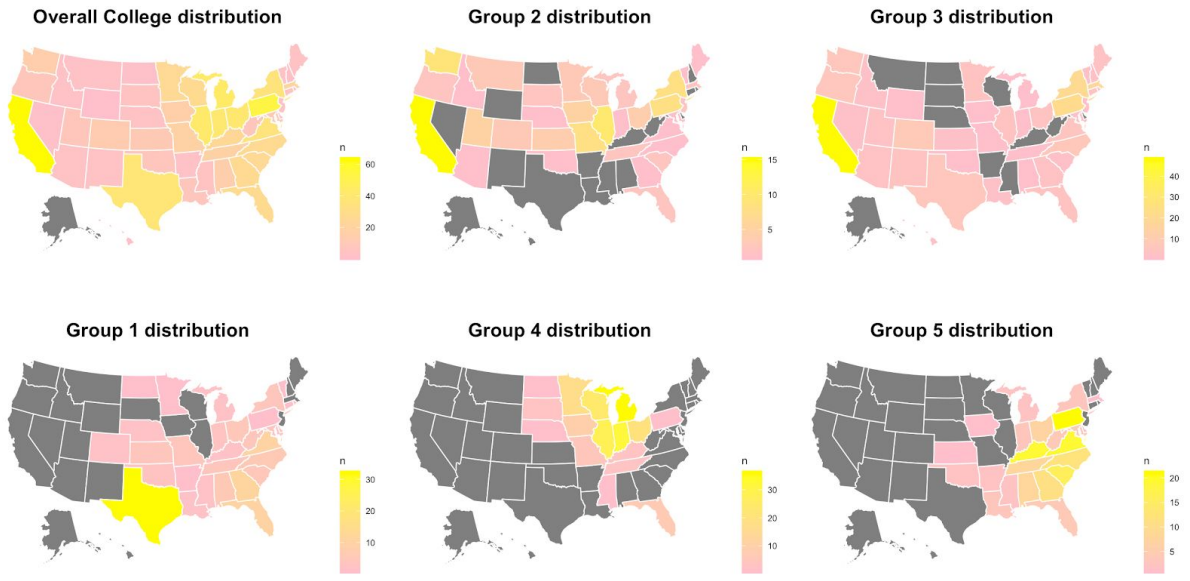

Figure 7, Geographical Distribution of Colleges

Overall, we conjecture that: for academically talented students, reputation is the first concern that outweighs school type and region; high school students in California, New York and Pennsylvania are more willing to consider schools far away from their hometown, compared with those in Texas and the Great Lakes region. Vice versa, colleges in California, New York and Pennsylvania tend to be more embracing for students nationwide and have an across-country charm.

Conclusions

In this manuscript, we start our study on students' preferences in the college decision-making process by using the Bradley-Terry model to create a preference ranking out of pairwise comparison data from Parchment, and then apply Newman's community detection algorithm on matchups between schools. The analyses suggest that universities possessing an advantage on Computer Science and Engineering are slightly preferred, which implies that such preference might be job-oriented. The results from community detection provides clusters of schools for students to refer to, e.g. looking at neighboring schools in the same group and understanding what other students with similar target schools would consider applying. The results also reveal that while high-achieving students would take reputation as first concern, typical students exhibit a tendency to consider schools not far from their residence. It is also suggested that students in California, New York and Pennsylvania are more willing to apply to faraway colleges, compared with students residing in Texas, Great Lakes region and other South Atlantic regions.

Reference

Bradley, Ralph, and Terry, Milton. “Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons.” *Biometrika*, vol. 39, no. 3/4, p. 324, 1952

Causeur, David, and François Husson. “A 2-Dimensional Extension of the Bradley–Terry Model for Paired Comparisons.” *Journal of Statistical Planning and Inference*, 23 July 2004.

Connie, Gan and Rahman, Abdul. “Exploring Key Factors Influencing University Choice.” *First International Digital Conference on Modern Business and Social Science*, December, 2018. doi: 10.13140/RG.2.2.16662.80962.

Efron, Bradley. *Bootstrap Methods: Another Look at the Jackknife*. Stanford University. Division of Biostatistics, 1977.

Hausman, Charles and Bradshaw, G. and Espinoza, Suzanne. “The College Decision-making of High Achieving Students.” *College and University*, vol.77, no.2, pp. 15–22, 2001.

Newman, Mark. “Finding Community Structure in Networks Using the Eigenvectors of Matrices.” *Physical Review E*, vol. 74, no. 3, pp. 36-104, 2006, doi:10.1103/physreve.74.036104.

Pearson, Egon . *'Student,' a Statistical Biography of William Sealy Gosset.* Oxford, UK: Clarendon Press, 1990

Usami, Satoshi. “Individual Differences Multidimensional Bradley-Terry Model Using Reversible Jump Markov Chain Monte Carlo Algorithm.” *Behaviormetrika*, vol. 37, no. 2, pp. 135–155, 2010. doi:10.2333/bhmk.37.135.

White, Scott, and Smyth, Padhraic. “A Spectral Clustering Approach To Finding Communities in Graphs.” *Proceedings of the 2005 SIAM International Conference on Data Mining*, 2005, doi:10.1137/1.9781611972757.25.

NACE Staff. “Top-Paid Advanced Degree Majors: Computer Science, Business.” *National Association of Colleges and Employers*. Accessed 17 April 2020.

“How U.S. News Calculated the 2020 Best Colleges Rankings.” *US News*. Accessed 17 April 2020.

“QS World University Ranking Methodology.” *QS Top Universities*. Accessed 17 April 2020.