2015 16th IEEE International Conference on Mobile Data Management

Exploring Institution-Driven Mobility

Which universities attract athletes from distant and diverse locales?

Clio Andris
Department of Geography
The Pennsylvania State University
University Park, PA, USA
clio@psu.edu

Zoe Andris
Department of Anthropology
Kenyon College
Gambier, OH, USA
andrisz@kenyon.edu

***Abstract*— Institutions are powerful forces in modern society—they affect human mobility, and are constantly pushing and pulling individuals across geographic space. Institution data is rarely including in studies of mobility, although we find these data to be a rich source of information on human mobility.**

We visit over 1000 U.S. university athletics webpages from 78 universities, 20 different types of sports, to produce a 90,000+ record dataset of U.S. and international student athletes and their hometowns. This new agent-based origin-destination data set has not been previously gathered into one file for public use.

For each university, we measure its "pull power", i.e. its ability to attract students from many different, and distant hometowns. From these data, we explore the mobility of students from different universities and find witch institutions are catalyzing human migration and movement to 'powerful' extents.

Keywords—migration; internet; data scraping; university; college; sports; ranking; social flows; social networks

I. Introduction

A. *Case Study: Human Mobility and University Choice*

Institutions such as the military, higher education, family, firms, ethnic and religious groups drive much of human mobility [1]. Universities spurn mobility in part by drawing migrants, i.e. students, from a wide and diverse set of locales. The wider and more diverse the student pool is, the more a university can be considered "powerful", in their ability to change the pattern of human settlement, as this school was chosen over intervening opportunities between the migrant's hometown and the school [2]. For instance, a small junior college may not appeal to students from outside of the state, as there are better opportunities closer to the student, perhaps in his or her own state. Conversely, a student may choose a large school with an international reputation rather than a closer university, as there are fewer academically-competitive options that are also near to one's home.

Previous studies in human mobility have utilized models of spatial locations, temporal movement, and social network ties to better predict movement of individuals. Through these explanatory models, human mobility can either be seen as a diffuse process or as a stochastic process that is time-varying and occurs around several fixed points [3]. Studies of human mobility are depending on study scope and geographic scale contexts, and in our experience, most often focused on mobility within a city. Recently, the field has benefited, or at least transformed, by an influx of GPS, social media, and cellular telephone network information that has allowed researchers to take a data-driven approach to understanding the size of human activity spaces, their movement dynamics and likelihoods of movement in geographic space [4]. As a result, these data are used for predicting epidemic spreads [5] and purported for use in urban planning. Interestingly, many of these datasets and research questions are approached through the lens of physics and mathematics inquiry—and thus, findings include distance decay and scaling models. Our approach is different in that is uses a GIS to examine the movement between a network of U.S. places.

B. *Problem: Current Measures of University Appeal*

The public relies on university ranking systems to develop their perception of college prestige and reputation [6]. In turn, universities often use their high rankings as a source of validation to draw highly qualified applicants. Such popular sites include, Forbes, Times Higher Education, QS, and U.S. News and World Report (USN), arguably the most popular source for college rankings in the U.S.. These ranking systems can be very helpful for students to search for attractive schools by their ranking, preview student body composition and caliber, and assess their own acceptance potential.

Yet, current methods of valuating universities may be problematic since rankings are determined in party by opinions [6,7]. Raters may be biased toward the institution at which they previously studied or those that have similar academic methods, approaches and values [8-10]. Departments also can receive higher ratings and evaluations when their institution holds a strong reputation [11]. Although additional "objective" factors were added, such as student-to-teacher ratio and average student SAT score, reputational scores still heavily weigh the index and rankings are assigned through a series of private opinions [11] instead of a transparent, data-driven process. A USN editor admitted to the possibilities of 'gaming' their ranking system [12]. Schools can use tactics such as admitting lower-scoring

978-1-4799-9972-9/15 $31.00 © 2015 IEEE
DOI 10.1109/MDM.2015.28

IEEE computer society

students in the spring semester, when scores aren't measured [12].

C. Solution: Measuring a Place by Geographic "Pull Power"

A metric that reflects the geographic diversity of a school's student body may help provide the public more information about a university. Some programs, such as San Francisco's Hastings College of Law, publically encourage students from 'underrepresented' geographic backgrounds to apply, akin to a call for geographic affirmative action [13]. "Pull power" can be quantified as the distances and diversity/variety of places that an institution can draw people from, as is said casually, from "far and wide". We believe an institution's pull power is an undervalued indicator of its prestige, and a university that can draw a mixture of cultures and landscapes may be of great value to its students.

Some schools can recruit students from many locales, while others tend to attract only in-state students. For example, in 2002, baseball rosters of 10 recruits showed that that Baylor [1] (TX) (ranked 16th in the nation in 2002) attracted its 11 recruits from Texas while Tennessee's unranked team attracted 10 students from 7 different states (NCAA [2] Division I Baseball Rankings, accessed 2013).

Currently, the public does not have access to information on how well a university can attract a variety of students. Although sometimes, a school will release its statistics about how many states and countries from which its student body hails.

Still, the diversity of connections should be used to explore places [14]. The use of large data sets, digital information, ICTs and the web can be a great help. For instance, though measures of telecommunications, diversity has shown that cities in the United Kingdom with diverse communication patterns tend to have higher socio-economic characteristics than those with insular communication [15].

With this in mind, we source online information to answer the following research questions: (a) How can we explore metrics of an institution's geographic pull power? (b) which universities exhibit strong pull power? Methodologically: How can this process be improved in the future, and what are the biases of our method? Contribution: Can this method be extended to larger studies—e.g. pull power of various competing tourist destinations? Does this experiment effectively start the conversation of characterizing the places where people flock: from 'far and wide' or from 'just around the corner'? Do we show how institutions induce human mobility?

II. Data Methodology & Description

While our goal is to obtain hometown information on each student in a university's student body, this information is sensitive and private. Thus, we use student athletes as a proxy for the student body. Issues with this approach are explored in the Discussion.

A. Acquiring Online Student Athlete Information

We source public college athletics team rosters from 78 (62 public, 16 private) U.S. schools. Due to the unavailability of centralized athlete data sets, we collect public records of online student athlete rosters, as published on individual university athletic websites. For each team, we retrieve the each team member's name, hometown, season (year), team (sport) and current university. We gather available information on over 20 sports every 4 years (to eliminate repeated athletes in the typical 4-year college tenure). We downloaded multiple years when possible and delete repeat individuals, unless he or she appears to have changed schools. The compiled dataset is cleaned by fixing errors in place spelling, standardizing place abbreviations, and attaching place names when only high school name was given.

B. School Selection Criteria

We select public schools from states that have (a) a flagship university: the state's "main" public campus [16] and a counterpart: (b) a non-flagship public institution often founded with an agrarian, engineering or specialized focus. We initially chose this method to pair each state's (a) and (b) schools to see which had greater pull power, while normalizing the distance to other states and big cities (since both a and b had similar geographies). We choose schools that have a research focus[3] and Division 1-A NCAA sports designation, as these have a high impact on both education and public notoriety.

We include 16 private schools for experimental comparison. This rough sample includes: Boston College, Boston U., Brigham Young, Brown, Carnegie Mellon, Columbia, Dartmouth, Fordham, Georgetown, Harvard, Kenyon, MIT, Providence, Villanova, Wake Forest and Yale. We are currently collecting data to expand this set of private and public schools to 125 institutions.

C. Data Descripton

Our dataset holds 93451 athletes with hometowns. The number of students gathered for each school depends on the years available, but range from 356 (Boise State) to 4093 (North Carolina) (median = 1121). Student athletes hail from United States: 84,345, followed by Canada: 2068, United Kingdom: 668, Australia: 369, Germany: 327, Sweden: 200, New Zealand: 184, France: 144, Brazil: 133, Mexico: 118, Spain: 117, South Africa: 115, Norway: 113, Netherlands: 111, Serbia & Montenegro: 101, Jamaica: 100, and 168 other nations with fewer than 100 athletes. In the United States,

[1] Universities such as "Baylor University" or "Kenyon College" are referred to by their short name, i.e. "Baylor" or "Kenyon". An exception is "Boston College". University is also abbreviated as U. at times.

[2] National Collegiate Athletics Association (NCAA), 2015. Division of Research. Accessed online Jun 5 at http://ncaa.org.

[3] See http://carnegieclassifications.iu.edu/ . Most schools are Research 1 (R1) institutions.

76,689 (82%) of student athletes hail from urban areas, which is comparable to the U.S. (81%), according to the U.S. Census. The top listed hometowns in the U.S. are San Diego, CA (676), Cincinnati, OH (546), Houston, TX (528), Charlotte, NC (428), Albuquerque, NM (403), Los Angeles, CA (397), Chicago, IL (394), Miami, FL (388), Phoenix, AZ (367), Dallas, TX (359), and Louisville, KY (352). However, suburbs were often listed as hometowns, and if students were gathered by metropolitan area, these statistics may change. 642 users do not list a hometown.

The top five most popular sports (by participant number) are football (17515 athletes), track and field (10837), swimming (8055), soccer (7624) and cross country (6031). Over half (55.6%) of athletes are listed as male. The number of athletes pulled from each school ranges depending on the number of rosters available. Years span from 1990-present, with emphasis on years 2012-2013, 2008-2009, 2004-2005, in order to capture recent athletes and reduce repetition.

D. *Pre-processing*

Hometowns are geolocated using the *Mapquest* Geolocation API. Universities are geolocated using longitude and latitudes provided by the U.S. Dept. of Education.[4] All GIS analysis is conducted in the ESRI ArcMap environment. Graphs and statistics are computed in the R statistical computing environment.

E. *Analytic Methods*

1) *Mean Center Movement*

For each university, we find the mean geographic center of its students, (i.e. the central feature of the point pattern of student hometowns). This mean represents the area that minimizes the average distance for the group of students. If this mean distance is far from the university, we can infer that the university has a significant pull power.

2) *Standard Deviation Ellipses*

A standard deviation ellipse characterizes a two-dimensional point pattern by finding the longest major axis that includes 1 standard deviation of distances from the point distribution in that linear direction. An orthogonal minor axis is set to encompass points in the perpendicular dimension within one standard deviation of the total distance distribution in that direction. An ellipse is drawn that connects the major and minor axes. This method is drawn from the ArcMap suite of tools. A larger, and longer ellipse signifies more pull power.

3) *Hometown Variety Statistics*

For each university, we find number of unique hometowns (by name) that students are from. This number is then divided by the number of possible student athletes, based on our dataset, to produce a *variety ratio*. A higher ratio may be an indicator of pull power. The distribution of students from each hometown (i.e. most students from a few places, vs. some students from all places) is characterized by the *average ratio* of students distributed to each hometown. A smaller average percent means that students have a wider distribution among the listed schools. In other words, in a rank-size distribution, the lower ranked hometowns would be marked with a fat tail. Finally, we find the *standard deviation ratio distribution*, where a smaller standard deviation is likely to be an indicator of the how concentrated values are around this average.

4) *Distance Profiles*

To explore and compare the "pull power" of universities we experiment with the following methods: We visualize the "Distance Profiles" of proximal sets or pairs of schools by learning which schools tend to draw students from a wide range of locations by plotting each university's density distribution of distances between each athlete's hometown and chosen university.

III. Results

A. *Geographic Measures*

1) *Mean Centers*

The difference between the mean center and the university location are highest for western states: California Berkeley (1992.8 km), New Mexico (1561.3), Idaho (1380.8), Oregon (1264.4) and Arizona (1222.8) (Fig. 1). Oregon State, Arizona State, Washington and Washington State closely follow. UC-San Diego has the lowest west coast difference for a school by a large margin at 347.7. The highest value for a private school is 1110.3 km for MIT, followed by Dartmouth (NH) and Brown (RI) with 1043 km and 1035 km differences, respectively. These values not only include difficulty pulling student from and beyond the sparse mountain states, but are also affected by a significant international population, including a number of football players from Hawaii and American Samoa.

The lowest pull power in this regard include North Carolina (75.8 km), North Carolina State (113.8), Wake Forest (NC) (126.7), Virginia (144.8) and Kentucky (200.6) (Fig. 2). This is not entirely an artifact of geographic location, as proximal schools such as Virginia Tech (845.6) and Louisville (989.7) do not have such constraints.

2) *Standard Deviation Ellipses*

The results of the standard deviation ellipses show that the ellipse major axis (km) generally correlates with ellipse area (square km) (Fig. 2). Western public schools California-Berkeley, Idaho and Washington State also lead in these pull power statistics, though private university Harvard has the 5th highest values, and primary university Boston U, ranks 10th (Fig. 2). Florida State exhibits a strong pull power compared to western schools in its ellipse area. Lowest pull schools in this regard are North Dakota State, a major outlier, and U. North Dakota, followed by Pittsburgh, and private Georgetown (D.C.). This graph provides an interesting way to compare the various pull power of co-located schools such as Boston U. and Boston College, the latter of which has a significantly stunted pull power when compared to the former. We find that Boston College seems to recruit more

[4] United States Department of Education, 2014. Office of Postsecondary Education, Database of Accredited Postsecondary Institutions and Programs. http://ope.ed.gov/accreditation/GetDownloadFile.aspx

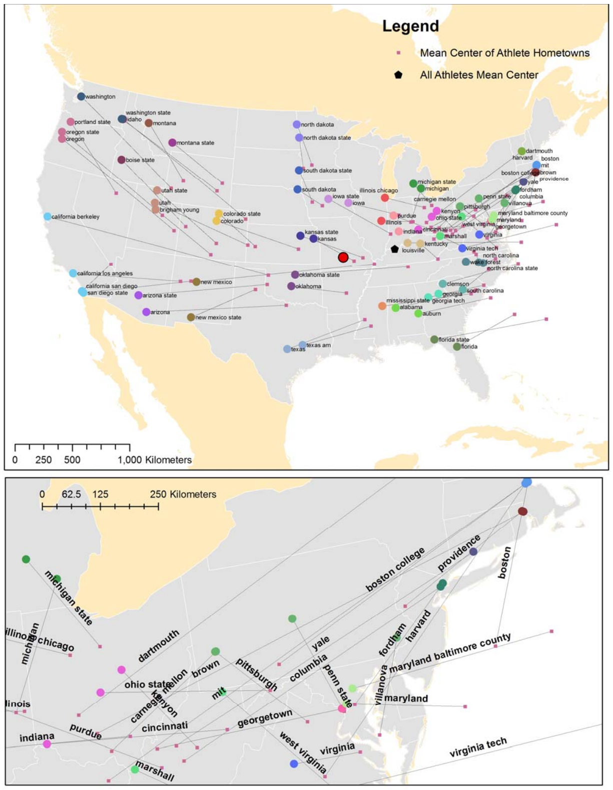

Fig. 1. Mean centers of university athletes compared with university location.

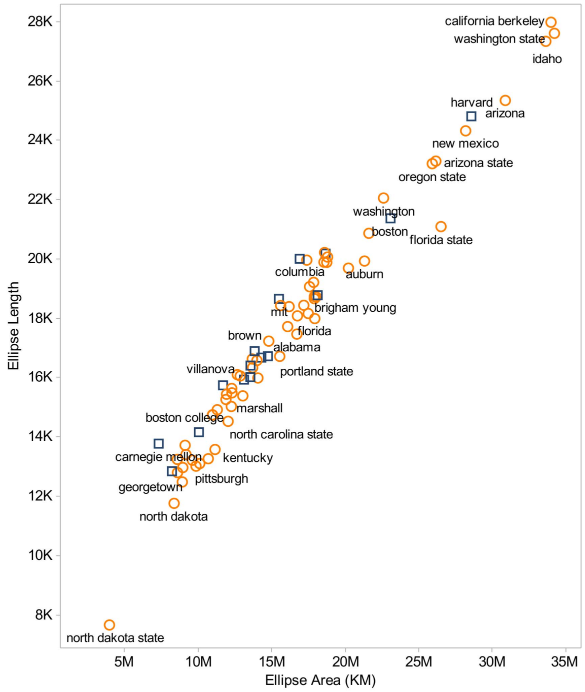

Fig. 2. Point patterns are characterized by ellipses. The area of the ellipse (in square km) is plotted against the major axis (length). Private schools are denoted as squares.

athletes from within the U.S., while Boston U.'s mean center is pulled to the east (Fig. 1).

B. Variety Statistics

Three schools have over 10% of their students coming from the same city. Each of these popular feeder cities are the cities where the university is located. At the U. New Mexico, 20.6% of students hail from Albuquerque, NM; U. Cincinnati holds 15.6% from Cincinnati and U. Louisville holds 11.21% from Louisville.

U. West Virginia has the highest variety of hometowns per student (0.80) (Table 1, showing top 20 ranked values) followed by a number of both private and public schools. The lowest variety of hometowns are in California: San Diego (0.32), Berkeley (0.38), San Diego State (0.41) and Los Angeles (0.42). Intermixed are Brigham Young (0.40), Washington (0.39) and North Carolina (0.38).

Although North Carolina does not show a great deal of unique hometowns, it has the lowest average ratio, indicating that students are spread over these different places (Table 1). Aside from Michigan, Ohio State and Penn State (which represent some of the nation's largest schools with 40,000+ students), these top pull power schools are mostly private schools (Table 1). The standard deviation of this ratio distribution is smallest for Boston U., followed by MIT, Dartmouth and Boston College (Table 1), although this value is slightly ambiguous.

Variety statistics might be best mitigated by the number of individual towns in the vicinity. For instance, New Mexico's population is concentrated in a few larger cities, while New Jersey has a sizable population in a number of small townships. However, theoretically, many towns instead of agglomerated places may still be a sign of more variety in hometowns, and thus, more diverse student bodies.

Lowest averages indicate that students are not from a high variety of places. These include New Mexico (0.89), Cincinnati (0.67), Montana (0.65), Portland State (0.58), Arizona (0.55) and Texas (0.54). Lowest standard deviations include Boise State (0.42), Utah State (0.40), Portland State (0.38), Clemson (0.35), Montana (0.34) and Virginia Tech (0.32)[5].

C. Distance Profiles

Distance profiles show private and public school differences by city, which affords consistence in the school's location and thus, its surrounding supply of students. In Boston, we find that Boston College (black) has the most local students, followed by Boston U. (red). These schools are in juxtaposition to MIT (blue) and Harvard (green), which draw students from a more variable location profile (profiles are truncated at 10,000 KM, but reach beyond this value) (Fig. 3).

In Washington, D.C., University of Maryland-Baltimore County (black) has a significant local spike, followed by Maryland (blue), which shows international activity (Fig. 3) reaching over 6000 km. Georgetown U. (red) has more pull

[5] Georgia is removed in this section due to many missing hometown values.

power with its varied trajectory, although it draws fewer international students. This trend is similar for others locales: In New York, we juxtapose Fordham (black) with Columbia (red). In this case, Fordham resembles a public school, after comparing it with public U. Pittsburgh (black) vs. private Carnegie Mellon (blue), which resembles private Columbia. We find these distance profiles to be very revealing about the nature of pull power—across the nation, and within Midwestern and international cities (as U. Pittsburgh exhibits).

TABLE I. Top 20 Schools in Geographic Variety and Distribution (Average & Standard Deviation)

Rank	School	Variety Ratio	School	Avg. Ratio	School	St. Dev. Ratio Distribution
1	West Virginia	0.8	North Carolina	0.063	**Boston**	0.069
2	**Providence**	0.75	**Brown**	0.078	**MIT**	0.088
3	**Villanova**	0.73	**Dartmouth**	0.08	**Dartmouth**	0.09
4	Clemson	0.71	**Boston**	0.081	**Boston College**	0.09
5	**Carnegie Mellon**	0.71	**Boston College**	0.088	**Brown**	0.096
6	**Kenyon**	0.68	**Columbia**	0.089	**Providence**	0.102
7	Boise State	0.67	**MIT**	0.09	**Fordham**	0.103
8	Virginia Tech	0.67	Michigan	0.104	**Columbia**	0.104
9	Penn State	0.66	Ohio State	0.108	**Harvard**	0.114
10	**Fordham**	0.65	**Yale**	0.112	**Georgetown**	0.116
11	Oklahoma State	0.64	**Georgetown**	0.112	**Yale**	0.118
12	Marshall	0.64	Penn State	0.112	**Villanova**	0.126
13	**Wake Forest**	0.62	**Harvard**	0.115	Penn State	0.135
14	**Boston**	0.62	**Brigham Young**	0.116	Michigan	0.196
15	Idaho	0.61	**Fordham**	0.125	Marshall	0.197
16	Louisville	0.61	Iowa	0.126	Iowa	0.2
17	**MIT**	0.6	Kentucky	0.131	North Carolina	0.205
18	**Harvard**	0.59	California Berkeley	0.131	Ohio State	0.206
19	Pittsburgh	0.59	Indiana	0.135	Maryland	0.209
20	Mississippi State	0.59	Maryland	0.135	**Kenyon**	0.211

IV. Discussion and Conclusion

There are a number of limitations to our results, primarily due to biases with our data set. First, the prestige of a certain athletic teams, i.e. its ranking against other teams and championship records, may skew the data so that students from these breeding grounds, or from a wider variety of hometowns, are drawn to certain schools because of the specific team's favorable reputation. Without the prominent team, they may not have been interested in the university. Also, coaches of teams may recruit students heavily from their own personal preferred regions.

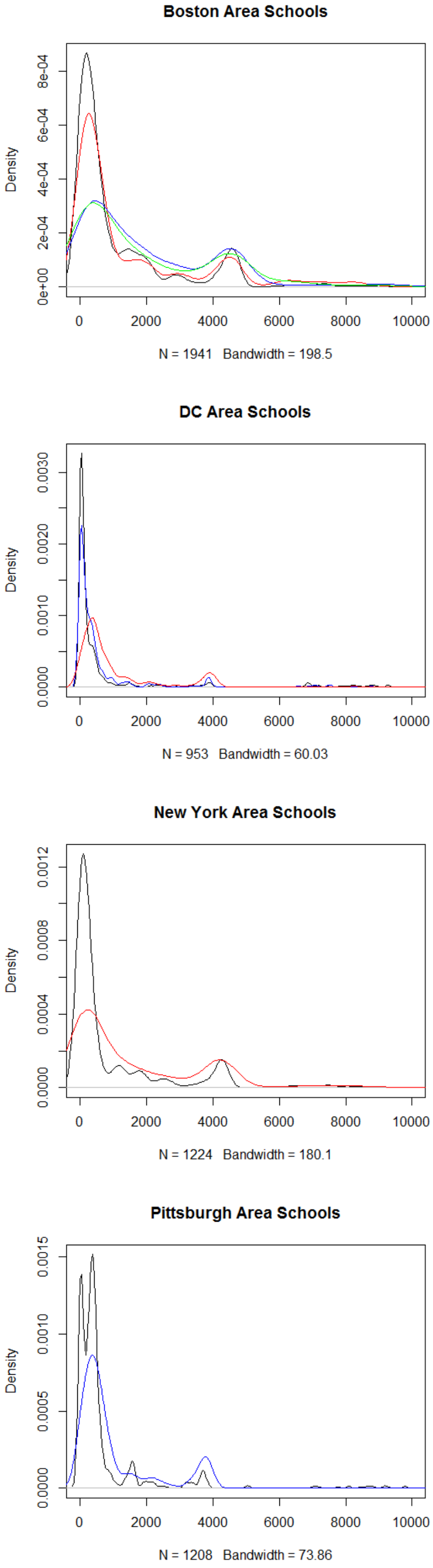

Fig. 4. Distance profiles for universities by city, legend is described in text.

A. *Sample Bias due to Athletic Culture*

There are a number of challenges involved with using athletic data. Foremost, it is not clear whether athletes are a good representation of a university's student body. Some geographies / high school athletic programs may be overrepresented in the athletic domains (called "breeding grounds" and places with hi "spatial 'production' rates") due to a disproportionate number of local students with strong athletic abilities in a particular sport [17, 18].

We find such "breeding grounds" for different sports. Many sports depend on climate (such as golf and ice hockey), but others seem to have a coast-bias (such as lacrosse), and an international draw (such as skiing, squash and tennis). Certain sports draw many international athletes: ice hockey 3,410/22,959 (12.9%), tennis 9,524/66,947 (12.4%), squash 310/3,144 (9.0%), and some very few: football 841/264,531 (0.32%) and wrestling 129/26,388 (0.48%) (NCAA 2014). In our sample 28.8% of ice hockey players, 35% of tennis players are international, 21% of squash players are international, and 1.4% of wrestlers and 1.2% of football players are international students.

B. *Future & Ongoing Work*

There are a number of improvements we plan to make in this analysis. First, we plan to complete the analyses with more schools in order to better understand the differences between private and public school's ability to draw students. Upon completion we also plan on publishing this dataset so that others can explore how human decisions tie places together. Next, we would like to explore other spatial statistical methods and quantifiable distributions of distance and geography. In further studies, we would like to have information on the hometowns of the student body at large, in order to thwart biases in representing a school by its student athletes.

C. *Conclusion*

Despite its exploratory nature, we believe this early exploration takes advantage of digital online data and provides insight for a methodology that helps us better understand how humans move in geographic space. At other geographic scales of human mobility, the measures explored here could be used to understand which institutions are most seminal in catalyzing migration and mobility—at what distances and with what variety. For instance, a new senior citizen living facility may create new mobility patterns for both the residents and visitors. How would this addition compare to a new movie theater or concert venue? Our case study benefits from this analysis: we find that measuring different facets of a school's pull power, as an additional factor alongside the university's rank. This information could be beneficial for both perspective students and urban, regional, and transportation planners. With these metrics we better understand mobility patterns catalyzed by universities, as well as the composition of the student body—for instance, that 20% of U. New Mexico athletes are from Albuquerque, NM. We believe there is great value to understanding an institution—whether a military base or a 'Chinatown'--by the patterns of mobility it drives.

Acknowledgment

We thank Ransom Hollister and Jason Gochanour for assistance with data collection. We thank Carlos Baez for helpful discussions.

References

[1] G. Bertocchi and C. Strozzi, "International migration and the role of institutions," *Pub. Choice*, vol. 137(1-2), pp. 81-102, October 2008.

[2] S. A. Stouffer, "Intervening opportunities: a theory relating mobility and distance," *Am Soc. Rev.*, vol. 5(6), pp. 845-867, December 1940.

[3] Cho, E., Myers, S. A., Leskovec, J. "Friendship and mobility: user movement in location-based social networks," Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1082-1090, August 2011.

[4] Song, C., Qu, Z., Blumm, N., and Barabási, A. L. "Limits of predictability in human mobility," *Science*, vol. 327(5968), pp. 1018-1021, 2010.

[5] Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., and Buckee, C. O. "Quantifying the impact of human mobility on malaria," *Science*, vol. 338(6104), pp. 267-270, 2012.

[6] R. Brooks, "Measuring university quality," *Rev. Higher Edu.*, vol. 29(1), pp 1-21, 2002.

[7] A. Machung, "Playing the ranking game," *Change*, vol. 30(4), pp. 12-16, 1998.

[8] A.M. Cartter, "An assessment of quality in graduate education," *Washington, DC: American Council on Education,* 1966.

[9] D.S. Webster, "Advantages and disadvantages of methods of assessing quality," *Change,* vol. 13, pp. 20-24, 1981.

[10] L.V Jones, G. Lindzey, and P.E. Coggeshall, "An assessment of research doctorate programs in the United States," Washington, DC: National Academy Press, 1982.

[11] N. Diamond, and H.D. Graham, "How should we rate research universities?" *Change*, vol. 32, pp. 20-33, 2000.

[12] M. Kutner, 2014. "How to Game the College Rankings" *Boston Magazine*, September 2014 [Online]. Available: http://www.bostonmagazine.com/news/article/2014/08/26/how-northeastern-gamed-the-college-rankings/. [Accessed Jan. 22, 2015].

[13] C. Andris, J. Wittenbach, and D. Cowen, "Support vector machine for spatial variation," *Trans. GIS*, vol. 17(1), pp. 41-61, February 2013.

[14] M. Batty and J. Cheshire, "Cities as flows, cities of flows," *Env. Plan B*, vol. 38(2), pp. 195-196, March 2011.

[15] N. Eagle, M. Macy and R. Claxton, "Network diversity and economic development," *Science*, vol. 328(5981), pp. 1029-1031, May 2010.

[16] S. Key, "Economics or education: The establishment of American land-grant universities," *J Higher Edu.*, vol. 67(2), pp. 196-220, March 1996.

[17] J. Rooney and R. Pillsbury, Atlas of American Sport, New York: MacMillan, 1992.

[18] J. Bale, "Human Geography and the study of sport," In; Handbook of Sports Studies. Ed. Jay Cakley, Eric Dunning. London: Sage, 2000.