ABSTRACT

Social Network Analysis in College Choice

Meiqing Ren, M.S. Eco.

Mentor: Van Pham, Ph.D.

This research paper employs Social Network Analysis to examine the effect of peer groups on the college application process, taking Baylor University as a case study. Degree centrality and eigenvector centrality are two centrality measures used as interested independent variables. Findings reveal that peer groups have significant effects on college choice in terms of high school and home neighborhood networks. Specifically, a one standard deviation increase in degree centrality implies about a 1.5% rise in application rates, which means highly connected students with a higher degree centrality are more likely to apply to Baylor University. This paper indicates that Baylor University has attracted applicants who are clustered by ZIP Codes, showing that if a student lives or studies in an area where lots of peers are also identified by Baylor University as potential recruits, he or she will be more likely to apply to Baylor. Thus, our study helps to broaden strategies for college recruitment by exploring the important role of social networks.

Social Network Analysis in College Choice

by

Meiqing Ren, B.Mgt.

A Thesis

Approved by the Department of Economics

_____

Charles North, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of

Master of Science in Economics

Approved by the Thesis Committee

_____

Van Pham, Ph.D., Chairperson

_____

James West, Ph.D.

_____

Scott Cunningham, Ph.D.

_____

Constanze Liaw, Ph.D.

Accepted by the Graduate School

May 2017

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES v
LIST OF TABLES vi
ACKNOWLEDGMENTS vii
CHAPTER ONE 1
Introduction 1
CHAPTER TWO 4
Literature Review 4
College Choice 4
Peer Effect 5
Social Network Analysis 6
CHAPTER THREE 9
Methodology 9
Population 9
Research Design 10
Sample Statistics 18
Variables 19
CHAPTER FOUR 21
Empirical Analysis 21
Empirical Models 21
Analytical Results 22
Robustness Check 30
CHAPTER FIVE 34
Conclusions and Discussions 34
Conclusions 34
Limitations 34
Recommendations for Future Studies 36
APPENDICES 37
BIBLIOGRAPHY 47

LIST OF FIGURES

Figure 2.1. Social Network Example……..7

Figure 3.1. Example Network……11

Figure 3.2. Social Network Visualization of a Sub-Sample……..17

Figure A.1. Geographic Area of ZIP's First Digit…….38

LIST OF TABLES

Table 3.1. Sociomatrice Format of the Example Network……11

Table 3.2. Edge list format of the Example Network……11

Table 3.3. Original Data Format Example……13

Table 3.4. Sociomatrix Data Format after Transformation……14

Table 3.5. Sociomatrix Data Format Using a New Tie Scale……14

Table 3.6. Sample and Population Comparison……19

Table 4.1. Regression Results of School Network……24

Table 4.2. Regression Results of Home Network.....25

Table 4.3. Relationship between Two Centralities in School Network……26

Table 4.4. Relationship between Two Centralities in Home Network……..26

Table 4.5. Regression of School Network……28

Table 4.6. Regression of Home Network……...29

Table 4.7. Robustness Check of School Network……..31

Table 4.8. Robustness Check of Home Network……...32

Table 4.9. Robustness Check Using the Population…….....…33

Table B.1. Data Dictionary……39

Table B.2. Summary Statistics of Centrality…….....……40

ACKNOWLEDGMENTS

I would like to thank Dr. Van Pham for his great directions and suggestions throughout this project. I would also like to thank the rest of my committee, Dr. James West, Dr. Scott Cunningham, and Dr. Constanze Liaw for their support and input. To Baylor admissions office and IRT office, thank you for giving me access to this valuable dataset and for helping me fix technical problems. And finally, I would like to thank my parents for supporting my choice of studying abroad and being a part of this excellent program.

CHAPTER ONE

Introduction

In the increasingly competitive environment of higher education, many universities have begun to employ sophisticated business methods to position themselves better in the market. Thus, their admissions offices are looking for broader strategies of recruiting high quality applicants. This study aims to use social network analysis to examine patterns in student interactions through high school and neighborhood networks. The significance of this study is that it offers helpful suggestions for college recruitment offices in identifying a potential pool of desirable students and in implementing new recruitment techniques.

For years, America's colleges and universities have boomed with more and more students. According to the latest national data from Nation Center for Education Statistics (NCES), an increasing number of students have been getting their high-school diplomas over the years. However, it doesn't mean more high school graduates have been going to college. As NCES reported, the percentage of all high-school graduates who immediately enrolled in college fell from 69 percent in 2008 to 66 percent in 2013, while in 2015 the percentage of high school graduates enrolled in college rose to 69.2 percent. Such a fluctuation of college enrollment aligns with the realities of shifting demographics and increasing competition in higher education. What's more, top and affluent students today have more college choices than ever before because of better technology, less expensive methods of communication as well as transportation, and richer financial support from

governments or institutions. Also, admissions recruiters are facing restricted resources, such as budget limits, as well as financial challenges from prospective students and parents. These issues collectively require admissions recruiters to explore different recruiting strategies since many of them are seeing that the usual admission procedures are no longer effective. They need to not only understand the motivations and mindsets of prospective students but also come up with new recruitment ideas under limited budgets.

This study's research question is how social networks through high school and neighborhood peer groups affect students' decisions of whether or not apply to a higher - education institution; Baylor University is used as a case study. Baylor University (BU), founded in 1845, is a private Baptist university in Waco, Texas, and has more than 16,000 students. In this paper, I use administrative data which was collected from high school students in the US to examine the importance of network influences on Baylor University's applications. My dataset includes students' contacts with Baylor (by email, phone, campus visits or other ways), academic performance, college enrollment decisions and other statistics. The main methodology in this paper is to build two social networks including high school and neighborhood network according to the distance between each respondent, which is estimated by ZIP Code information. In this way a social network dataset is created and corresponding tools can be employed to explore whether peer networks are associated with an individual's college choice.

The remainder of the paper is organized as follows. Chapter 2 provides a review of the literature associated with college choice, peer effect and social network analysis. Chapter 3 describes the methodology, including population, research design, sample

statistics, and variables. Chapter 4 presents empirical analysis. Chapter 5 makes conclusions, discusses limitations, and considers future research.

CHAPTER TWO

Literature Review

College Choice

As student recruitment has become increasingly important, numerous studies have examined the college choice process in an attempt to identify factors influencing student decision making. Many scholars have employed economic and sociologic theoretical frameworks to examine factors of college choice (Hearn, 1984; Jackson, 1978; Tierney, 1983; Somers, Haines, & Keene; 2006), which include three kinds of models: (a) economic models, which means students rely on careful cost-benefit analyses (b) status-attainment models, which assumes a utilitarian decision-making process that students go through in choosing a college and (c) combined models, which combine the former two.

Much of the literature on this topic that may affect college choice discusses how several different factors play a role in the college choice process of students. Some of this literature focuses on matching student features to institutional features (Zemsky & Odell, 1983), while other literature studies what characteristics of institutions are needed affect student decisions (Martin & Dixon, 1991; Martin, 2006; Paulsen, 1990). In the early 1990s, major factors that motivated students to enroll in college were divided in four clusters: academic program, location and climate, cost, and influences of important people (Johnson, Stewart, and Eberly, 1993; Martin and Dixon, 1991; Sevier, 1994). With the higher-education environment becoming increasingly competitive, more factors have been considered and analyzed. For example, “push and pull” factors are identified as

influencing students' decision making processes around the world since studying abroad is a popular way to experience different cultures (Zimmerman, 2000). In general, the areas that might influence college choice currently fall into five kinds of categories: family effect; influence of peers, institutional characteristics; institutional contact; and institutional fit (Furukawa, 2011). Peer influence is what this study focuses on.

Peer Effect

As for the definition of a peer group, it consists of people who have similar backgrounds, interests, or social status. Teenagers' peer groups consist of individuals of similar ages, though teenagers can belong to various peer networks, such as friends, classmates, and teammates in different social situations (McNeal, 1995). A peer group is also defined as a "collection of individuals with whom the individual identifies and affiliates and from whom the individual seeks acceptance or approval" (Astin, 1993)

There is much literature that documents the significance of peer influences on adolescent choices. Peer effect refers to both student peer groups (Kealy & Rockel,1987; Kern, 2000) and individual groups to which students select to get along with (Johnson & Stewart, 1991; Burleson, 2010). Some literature looks at peer influence on students' decisions to attend college. For instance, individuals with more classmates who have matching college preferences are more likely to enroll in these colleges (Fletcher, 2010). Others focus on how a peer group influences students' college achievements measured by grade point average (GPA), persistence, etc. (Fletcher & Tienda, 2009). For example, peer choices and characteristics have been demonstrated to be significant in predicting students' performances during middle school (Summers & Wolfe 1977, McEwan 2003, Lavy & Schlosser 2007), high school (Ding & Lehrer 2007), as well as their achievement

in college (Sacerdote 2001, Zimmerman 2003). Furthermore, peer effects during college persist at a diminishing rate into the sophomore, junior, and senior years, indicating that social network peer effects may have long lasting effects on students' academic achievements (Carrell, Fullerton & West, 2008). While those literature believes in the effects of peer influence on students' college choice as well as academic performance, other researches argue that there is no relationship between choice decision and peer effect (Hossler, Braxton, & Coopersmith, 1989).

Social Network Analysis

The Social network analysis (SNA) is a particular methodology which has its own type of data collection, statistical analysis, and results visualization. The nodes in the network are the people and groups while the links indicates relationships or flows between the nodes. It enables scholars, practitioners, and educators to study how "people are located or 'embedded' in the overall network" (Hanneman, 2001), and it concentrates on the nature and results of ties between nodes which are always individuals or groups (Scott, 2000; Wasserman & Faust, 1994).

Social network analysis can help us see where an actor is located in the network and how this network has been structured. As depicted in Figure 2.1, Person A and Person B have different positions in the network; Person A occupies a more compact network location than Person B. Person A's social construction may result in a social context with higher dependence and closer affinity, while Person B's network generates broader access to information with greater diversity.

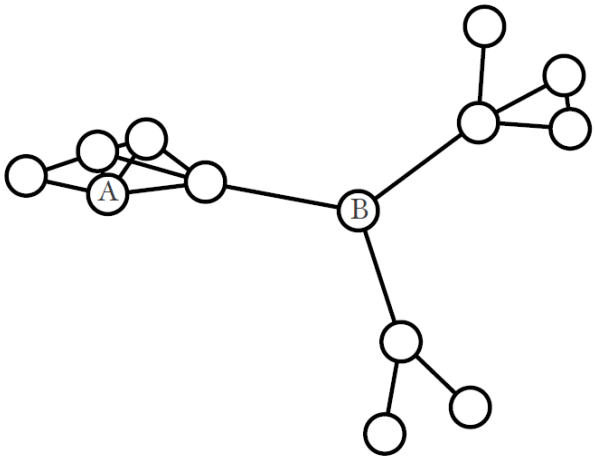

Figure 2.1. Social Network Example

Social network topics can vary widely, such as how people take advantage of their network to find jobs (Granovetter, 1973) and how companies use alliances to adapt to rapid economic changes (Stark & Vedres, 2005). Many studies look at the network through Social Media. For example, Ruane and Koku (2014) demonstrated that online peer mentoring sites offered support for interactions between first-year and third-year undergraduate education students. Some literature merely uses non-empirical ways to describe how social media works through the admission recruiting process in colleges and universities (Anton, 2006; Davis, Deil-Amen, Rios-Aguilar & Canche, 2012; Johnson, 2011; Kessler, 2011; Lavrusik, 2009), while some literature presents an empirical analysis of social media's value as a tool in college admissions (Ferguson, 2010; Dooney, 2014). Network analysis is also employed in education research. For example, Kapucu, Yuldashev, and Demiroz (2010) used SNA methods and tools to evaluate student interaction in an MPA class at the University of Central Florida, and different kinds of centrality such as degree, closeness, and betweenness centrality are employed to identify characteristics of student friendships and advice networks. Maroulis and Gomez

(2008) remind that it is necessary to consider the network structure of students' relationships when examining the influence of peers. However, compared with the many studies focusing on peer influence on college choice, there are few studies employing SNA tools and methodologies to examine the network effect on college choice.

CHAPTER THREE

Methodology

Population

I use applicant-level cross-section data to explore my research questions empirically. It is administrative data acquired by the Baylor Admissions Office; thus, it is not publicly available. Baylor University identified the students in the dataset as potential recruits and reached out to them several times beginning in 2001. Baylor collected this dataset in a rolling process, and more and more high school graduates have been added since then. For this study, I chose to focus on 122,967 students who graduated from high school in 2014. They are from all of the states in the US, and 54% are from Texas. Of the total students, only 26% applied to Baylor University, while 74% did not. Within those who applied, only 11% finally enrolled in Baylor. Of the 122,967 students, 43% are male, 3.25% have Baylor alumni in family, and 8.4% are Baptist.

To determine whether or not Baylor University can represent some type of college or university in the US, let us look at basic statistics and facts of higher education first. According to NCES, there was a total of 4,695 higher education institutions across the United States for school year 2015-2016, including 3,023 four-year and 1,672 two-year public and private (profit and not-for profit) degree-granting institutions. In 2015-16, the approximate number of undergraduate enrollments in universities across the United States was about 22.8 million, within which three fourths were students who attended public schools. In 2016, Baylor University, the four-year non-profit private school, had a

total enrollment of 16,959 (14,348 undergraduate and 2,611 graduate/professional students). This number places Baylor in the "large" category of college size which includes schools that have more than 15,000 students. And according to the 2017 Best Christian Colleges ranking from NICHE, Baylor University ranks 15th based on key statistics and student reviews using data from the U.S. Department of Education. Based on this limited information, Baylor University potentially represents large and high quality Christian private schools across the United States. However, it is still hard to say whether our findings from Baylor University are able to reflect common rules between social networks and college choice.

Research Design

The goal of this research is to use social network analysis in order to offer useful recommendations to the Baylor Admissions Office; in this way, Baylor will be able to broaden its strategies for enhancing the number and quality of applicants. The original data in this study is stored in rectangular data structures, in which rows are used to depict observations (students), and columns represent individual variables. This is a common way to store data for many types of data analysis. However, network analysis requires a different type of data storage because it needs to explore more complicated relational structures. Consider the following simple example of an undirected network in Figure 3.1. Sociomatrices (as in Table 3.1) and Edge-Lists (as in Table 3.2) are two ways that computers can recognize and operate the underlying network data.

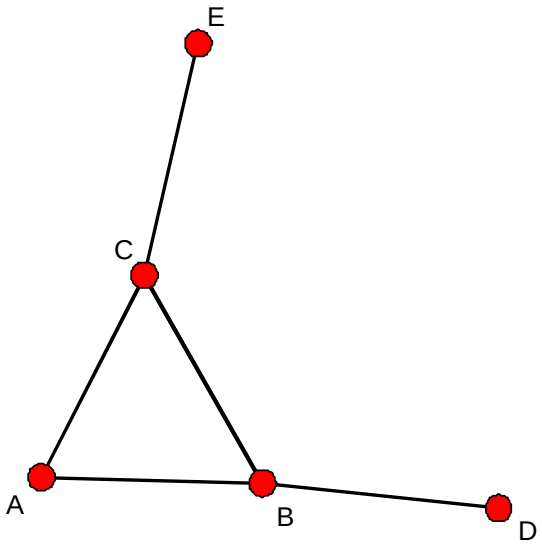

Figure 3.1. Example Network.

Table 3.1. Sociomatrice Format of the Example Network

Nodes	A	B	C	D	E
A	0	1	1	0	0
B	1	0	1	1	0
C	1	1	0	0	1
D	0	1	0	0	0
E	0	0	1	0	0

Table 3.2. Edge list format of the Example Network

From	To
A	B
A	C
B	C
B	D
C	E

This study uses the sociomatrix format, in which each cell indicates if the actors are related (1, 0) or the extent of the relationship. To transform the original dataset to a network adjacency matrix, I choose software Matlab (matrix laboratory), which is

designed for mathematical computation and analysis, as my transformation tool. In order to examine respondents' interactions within their high school networks or permanent home neighborhood networks, ZIP Code information can be used to reflect the distance between each respondent, and networks can be built by the scope of a certain distance from each student. However, Google Maps or Bing Maps, two popular tools for distance calculation, have some serious quota limitations. So I tried a new way of depicting a network: code different tie values in network matrices according to zip code digits which show how near two students are. That means if two respondents have the same high school ZIP Code or home ZIP Code, they have a closer relationship and thus get a larger tie value in the relationship matrix. But if they only have three or four digits of the same ZIP Code, their tie value in the sociomatrix is smaller. This method avoids limitations of distance calculations and makes the process of data transformation easier. This methodology is based on the story of ZIP Code.

A ZIP Code, which is a five digit number, identifies a specific geographic area and has been used by the United States Postal Service (USPS) since 1963. There are about 43,000 ZIP Codes in the United States. The first digit (0-9) of a ZIP Code represents a common area of the country with numbers starting lower in the east and increasing when you shift to the west (as shown in Appendix A). For example 0 designates Maine while 9 represents California. The next two digits are on behalf of one of the 450+ Sectional Center Facilities (SCFs) in America. A SCF (Sectional Center Facility) is a postal facility which works as the distribution and processing center for post offices at a specific area, designated by the first three digits of those post offices' ZIP Codes.

Accordingly, there is a certain rationality in believing that two respondents who have the same ZIP Code or at least the same first three or four ZIP Code digits have some kind of social interactions since they are relatively close together.

The methodology this study adopts in building a relationship adjacent matrix is explained as follows. Take the following respondent information as an example. Students A and D have the same 5-digit ZIP Code, so their corresponding tie value in the matrix is 3, which numerically represents the closest path distance and the strongest connection between two nodes. Students A and C have the same four ZIP Code digits, so their tie value which refers to their degree of closeness is coded 2. And the tie value of pairs (A, B) is 1 since only the first three digits are the same, while the tie value of pairs (A, E) is 0, indicating the farthest path distance and the weakest relationship between two nodes. Thus, the closer the students are to each other, the higher the tie value will be. This successful transforms my original data format like in Table 3.3 to the sociomatrix type of data storage, as in Table 3.4, that network analysis required.

In this way, our relationship matrix, which is symmetric and undirected, is accessible. However, Matlab has limitations for creating and exporting big matrices, so the maximum number of nodes in my networks is 6000. Finally, I built separate networks based on sample students' high school ZIP Codes and permanent address ZIP Codes and created two 6000×6000 adjacency matrices using the observations as network nodes. To do a robustness check in SNA, I further employed another kind of tie value scale while building matrices. So instead of the tie value set (3, 2, 1, 0) in which 3 refers to the closest relationship while 0 means the weakest connection, another set with different scales (1, 2/3, 1/3, 0) is also adopted, as in Table 3.5.

Table 3.3. Original Data Format Example.

Student	ZIP Code (home or high school)	
A	76706	Waco, TX
B	76710	Waco, TX
C	76702	Waco, TX
D	76706	Waco, TX
E	02215	Boston, MA

Table 3.4. Sociomatrix Data Format after Transformation.

Student	A	B	C	D	E
A	0	1	2	3	0
B	1	0	1	1	0
C	2	1	0	2	0
D	3	1	2	0	0
E	0	0	0	0	0

Table 3.5. Sociomatrix Data Format Using a New Tie Scale.

Student	A	B	C	D	E
A	0	1/3	2/3	1	0
B	1/3	0	1/3	1/3	0
C	2/3	1/3	0	2/3	0
D	1	1/3	2/3	0	0
E	0	0	0	0	0

As in any study of Social Network Analysis, we face many questions regarding how to define and evaluate the networks. In this paper, I begin by considering two questions.

The first question is whether the networks are directed or undirected. Directed networks contain only directed edges and undirected networks contain only undirected edges. In our case, two students are assumed to have some kind of connection if they have the same or similar ZIP codes, so we do not care about the direction of ties in our sociomatrix and we also do not have such information. Thus, our matrix is symmetric and undirected.

The other question is: what types of network measures in SNA are appropriate in our case? By examining the location of nodes, we are able to evaluate the prominence of those members. An actor is prominent if the actor is visible to the other members through his or her ties in the network (Knoke & Burt 1983). In fact, there are dozens of common ways available to network analysts to evaluate the prominence of network members. For our undirected networks, we will look at four types of centrality including degree, closeness, betweenness, and eigenvector centrality.

Degree centrality, which is an important measurement in our study, reflects the number of ties attached to the node. Degree centrality is based on the notion that a node that has more direct ties is more prominent than nodes with fewer or no ties since it has multiple alternative ways and resources to reach goals. Directed network defines two separate measures of degree centrality including in-degree centrality, which indicates the number of links coming in to an actor, and out-degree centrality which indicates the number of links going out from the actor. However, our sociomatrix is undirected, so

degree centrality is used to represent the number of ties that a node has regardless of directions. Thus, the degree centrality of each node is equal to the sum of its row in our matrix.

Closeness centrality refers to the number of links the node takes to reach everyone else in the network. It is defined as the inverse of farness, which in turn, is the sum of distances to all other nodes. The formula of closeness centrality is as follows.

$$C_c(i) = \left[\sum_{j=1}^{N} d(i,j)\right]^{-1}$$

Betweenness centrality shows the extent to which an actor lies on a path between other actors and thus controls information flow. It refers to how many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops. Betweenness is defined as follows, in which $g_{jk}(i)$ is the number of shortest paths connecting j and k passing through i, and g_{jk} is total number of shortest paths linking j and k.

$$C_B(i) = \sum_{j \neq k} g_{jk}\,(i)/g_{jk}$$

However, closeness centrality and betweenness centrality are inapplicable in our case since our relationship matrices are disconnected. Figure 3.2 displays an example of my social network visualization, in which the nodes are 12 students named from letter A to letter L with their corresponding ZIP Codes. The width of paths between each node, which is according to the set of tie values, represents the closeness of relationships between them. We can see that our whole network consists of dozens of non-cross cliques, and none of the nodes are necessary as conduits to pass information to other nodes.

Because the distance between nodes in disconnected components of a network is infinite, which means zero edge weights can produce an infinite number of equal length paths between pairs of nodes, so two measures of closeness centrality and betweenness centrality cannot be applied to networks with disconnected components (Opsahl, 2010; Wasserman & Faust, 1994).

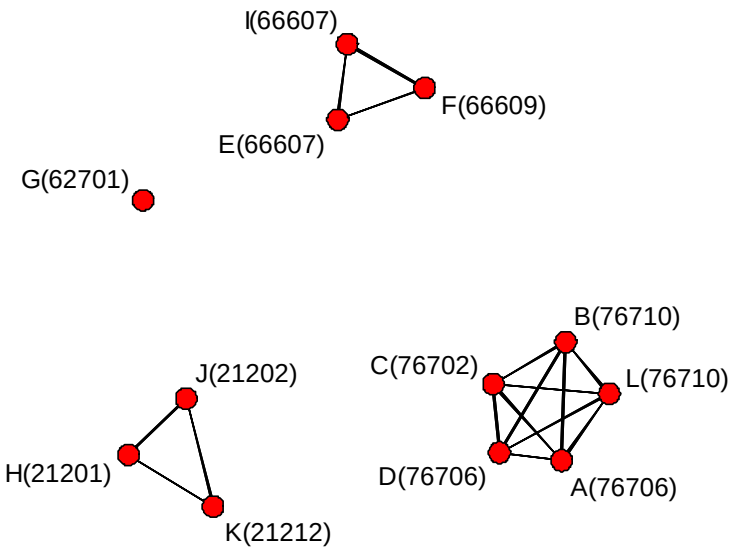

Figure 3.2. Social Network Visualization of a Sub-Sample

Finally, eigenvector centrality, a natural extension of degree centrality, reads that a node is of great importance if it is linked to other important nodes. The assumption is that each node's centrality is the sum of the centrality values of the nodes that it is linked to. Thus, a node having many links does not necessarily possess a high eigenvector centrality since a central node should be the one connected to high-scoring nodes. Furthermore, a node which has a high eigenvector centrality is not necessarily highly linked since it may possess few but highly important linkers. The eigenvector centrality

X_i of node i is given by the following formula. Let $A=(a_{i,j})$ be the adjacency matrix of a graph.

$$x_i = \frac{1}{\lambda} \sum_k a_{k,i}\, x_k$$

$$\lambda x = xA$$

Hence, the centrality vector x is the left-hand eigenvector of the adjacency matrix A associated with the largest positive eigenvalue λ. Thus, the main idea is to calculate the eigenvector of the largest positive eigenvalue as a measure of centrality.

Essentially, the degree centrality and eigenvector centrality are two parameters employed in this paper in order to explore the relationship between centralities and application decisions.

Sample Statistics

The dataset consists of a large population of approximately 122,967 students who graduated from high school in 2014. As I mentioned before, Matlab has limitations of transforming and exporting big matrices, so I randomly chose 6,000 students, approximately the maximum number acceptable by Matlab, as my sample for following regression analysis. Table 3.6 shows basic summary statistics of students from the sample and the population respectively. The similar distribution in each variable appears to show that our sample, which was obtained by random sampling, has similar measurable characteristics to our population; thus, our sample is representative of the population.

Table 3.6. Sample and Population Comparison.

Categories	Sample Mean*	Population Mean*
Applied-Yes	25.0%	25.9%
Applied-No	75.0%	74.1%
Enrolled-Yes	3.0%	2.9%
Enrolled-No	97.0%	97.1%
Gender-Male	43.7%	43.0%
Gender-Female	55.4%	56.1%
Ethnicity-White	42.9%	32.1%
Ethnicity-Hispanic	4.8%	4.9%
Ethnicity-Asian	4.7%	5.0%
Ethnicity-Black	9.5%	9.1%
Ethnicity-Amerind	1.8%	1.7%
Ethnicity-Pacisle	0.2%	0.3%
High school type-C	10.0%	10.2%
High school type-Public	85.1%	84.9%
High school type-Other	3.0%	2.7%
High school percentile- Min-25	0	0
High school percentile- 25-50	10.1%	10.2%
High school percentile- 50-90	8.0%	8.5%
High school percentile- 90-Max	2.2%	2.1%
Contact times	2.4	2.4
Total contact people	2.6	2.5

*There are some missing values, so the total percentage for each category may not be 100.

Variables

The primary independent variables are the degree centrality and eigenvector centrality of each student in two kinds of networks including high school and home neighborhood networks. Other independent variables acting as control variables contain gender, high school type, ethnicity, total contacts (the number of contact people between Baylor and the student), and high school percentile (student's ranked academic performance during high school). Contact times (how many times the student connects with Baylor via

email, phone, or campus visit) is another independent variable, and I regard it as an intervening variable since it can be controlled by not only students but Baylor University. Our dependent variable is whether or not the student applied to Baylor University. More details about the meaning of variables are available in Table B.1. So the main point is to explore whether social networks have significant effects on college choice by checking the coefficients of two centralities.

CHAPTER FOUR

Empirical Analysis

Empirical Models

Ordinary least squares (OLS) and logistic regression are two empirical models employed in this paper. When used with a binary response variable, OLS regression is known as a linear probability model and can be adopted as a way to assess conditional probabilities. It is most easily understood visually, and researchers always regard it as a benchmark for comparison. However, this linear probability model may lead to some problems. For example, the distribution of the error term in this case is not normally distributed but binomial. It indicates that not only our traditional t-tests for individual significance but F-tests for overall significance do not work here. In addition, using OLS method to estimate a model with a dummy dependent variable might perform badly if the true model is nonlinear and has a heteroskedasticity problem. It may also lead to predictions outside the range of 0 and 1. The Logit and Probit models fit my exploration of nonlinear estimation with a dichotomous dependent variable and solve the potential problems above, though the interpretation of coefficients is not straightforward. Both Logit and Probit methods will yield similar (though not identical) estimations. Beck (2011) believes that Logit not only outperforms with more uncommonly generated data than OLS but also appears to outperform regression for assessing binary (exogenous) treatment effects. So I chose to use Logit regression and OLS to explore my research questions.

In order to collect and analyze the data, our research method includes the use of three programming languages; Matlab, R, and STATA. As I mentioned before, Matlab can be used to create our adjacency matrices. R, a programming language and software environment for statistical computing and graphics, has several advantages as an ideal platform for developing and conducting network analyses. For instance, the R system includes dozes of packages that are designed to accomplish specific network analytic tasks. Compared with UCINET, a software package for the analysis of social network data, R is able to conduct larger-scale network analyses. R and STATA are both able to do OLS and Logit regressions and export results nicely, and in this paper STATA is employed as our regression tool.

Analytical Results

There is a significant relationship between our network centrality and the student's college choice (whether or not the student applied to Baylor). Previous research has produced strong evidence on the influence of peers on student college decisions. This study employs SNA tools and methodologies, and its findings align with previous research that found peer groups are indeed related to college choice. Degree centrality and eigenvector centrality are two centrality measures used as interested independent variables. Summary statistics of degree centrality and eigenvector centrality are displayed in Table B.2. I conducted a multiple regression to examine the relationship between the dependent factor and the independent variables as potential predictors, and two types of networks including high school and home neighborhood networks were analyzed individually.

Table 4.1 and Table 4.2 present the results of multiple regressions analyzing high school and home neighborhood network effects via OLS and Logit models. In this case, I used the tie value set (3, 2, 1, 0) to build my relationship matrix in which 3 refers to the closest relationship while 0 means the weakest connection. The results illustrate that coefficients of degree centrality and eigenvector centrality are both significantly positive when we look at single regressions, which means these two centralities have positive relationships with a student's application decision. However, when it comes to the results of multiple regressions, the coefficients of eigenvector centrality in both high school and home networks become negative while those of degree centrality are still positive. This situation leads us to think about the multicollinearity problem. One way to measure multicollinearity is the variance inflation factor (VIF), which assesses how much the variance of an estimated regression coefficient increases if predictors are correlated. If no factors are correlated, the VIFs will all be 1. The output below in Table 4.1 and Table 4.2 shows that the VIFs for degree centrality and eigenvector centrality in OLS3 models are about 1.7, which indicates some correlation though not enough to be overly concerned about. However, Valente, Coronges, Lakon, and Costenbader (2008) show that eigenvector centrality and degree centrality are highly interrelated because both measures are symmetrized and rely, to some extent, on direct connections. So we can expect these two different measures to behave similarly in statistical analyses. Table 4.3 and Table 4.4 show a robust relationship between two centralities in both types of networks. Considering my concerns that uncertain reasons may lead the coefficients of eigenvector centrality to fluctuate, I chose to exclude the eigenvector centrality as my independent variable. It is also reasonable to believe that degree centrality can tell us a similar story.

Table 4.1. Regression Results of School Network

VARIABLES	(1) OLS1	(2) OLS2	(3) OLS3	(4) OLS4	(5) Logit1	(6) Logit 2	(7) Logit3	(8) Logit4
schooldeg123	0.000590***		0.000687***	0.000124***	0.00296***		0.00352***	0.00251***
	(3.92e-05)		(4.79e-05)	(2.46e-05)	(0.000205)		(0.000252)	(0.000536)
schoolegvc123		2.522***	-1.887***	-0.533**		11.94**	-9.835***	-11.47**
		(0.445)	(0.534)	(0.269)		(2.142)	(2.624)	(5.768)
gender				-0.0236***				-0.489***
				(0.00559)				(0.133)
white				0.0600***				1.688***
				(0.00614)				(0.188)
hispanic				0.214***				4.979***
				(0.0139)				(0.386)
asian				0.0969***				2.237***
				(0.0136)				(0.289)
black				0.147***				2.708***
				(0.0101)				(0.214)
amerind				0.0909***				2.499***
				(0.0209)				(0.439)
pacisle				0.0191				
				(0.0647)				(omitted)
stype_c				-0.00246				0.163
				(0.0212)				(0.692)
stype_public				0.0156				0.520
				(0.0196)				(0.649)
stype_other				-0.00134				0.245
				(0.0258)				(0.790)
hs25_50				0.761***				7.114***
				(0.0132)				(0.775)
hs50_90				0.762***				
				(0.0137)				(omitted)
hs90_max				0.758***				
				(0.0214)				(omitted)
total_contacts				-0.00679**				-0.299**
				(0.00276)				(0.135)
contacttimes				0.0265***				0.677***
				(0.00406)				(0.152)
Constant	0.170***	0.242***	0.163***	0.00624	-1.540***	-1.143*	-1.589***	-4.944***
	(0.00762)	(0.00574)	(0.00789)	(0.0220)	(0.0448)	(0.0311)	(0.0471)	(0.702)
Observations	6,000	6,000	6,000	5,943	6,000	6,000	6,000	5,336
R-squared	0.036	0.005	0.038	0.762				

* Significant at the 0.10 level, ** Significant at the 0.05 level, *** Significant at the 0.01 level. Standard errors are in parentheses. The dependent variable is a binary variable indicating whether or not the student applied to Baylor University.

Table 4.2. Regression Results of Home Network

VARIABLE	(1) OLS1	(2) OLS2	(3) OLS3	(4) OLS4	(5) Logit1	(6) Logit 2	(7) Logit3	(8) Logit4
homedeg123	0.000612***		0.000748***	0.000163***	0.00305***		0.00378***	0.00309***
	(4.05e-05)		(5.00e-05)	(2.57e-05)	(0.000210)		(0.000263)	(0.000543)
homeegvc123		2.268***	-2.485***	-0.873***		10.83**	-12.42***	-14.96***
		(0.445)	(0.541)	(0.273)		(2.156)	(2.666)	(5.776)
gender				-0.0239***				-0.503***
				(0.00558)				(0.133)
white				0.0603***				1.704***
				(0.00614)				(0.189)
hispanic				0.212***				4.957***
				(0.0138)				(0.387)
asian				0.0968***				2.223***
				(0.0136)				(0.289)
black				0.146***				2.704***
				(0.0101)				(0.215)
amerind				0.0900***				2.473***
				(0.0208)				(0.440)
pacisle				0.0229				
				(0.0646)				(omitted)
stype_c				0.00435				0.287
				(0.0212)				(0.688)
stype_public				0.0215				0.630
				(0.0195)				(0.645)
stype_other				0.00323				0.304
				(0.0257)				(0.785)
hs25_50				0.762***				7.133***
				(0.0132)				(0.772)
hs50_90				0.763***				
				(0.0137)				(omitted)
hs90_max				0.758***				
				(0.0214)				(omitted)
total_contact				-0.00682**				-0.299**
				(0.00276)				(0.137)
contacttimes				0.0264***				0.674***
				(0.00405)				(0.155)
Constant	0.173***	0.243***	0.163***	-0.00132	-1.524***	-1.138**	-1.584***	-5.086***
	(0.00749)	(0.00574)	(0.00775)	(0.0221)	(0.0439)	(0.0310)	(0.0462)	(0.701)
Observations	6,000	6,000	6,000	5,943	6,000	6,000	6,000	5,336
R-squared	0.037	0.004	0.040	0.763				

* Significant at the 0.10 level, ** Significant at the 0.05 level, *** Significant at the 0.01 level. Standard errors are in parentheses. The dependent variable is a binary variable indicating whether or not the student applied to Baylor University.

Table 4.3. Relationship between Two Centralities in School Network

VARIABLES	(1) OLS
schoolegvc123	6,413***
	(118.0)
Constant	115.1***
	(1.523)
Observations	6,000
R-squared	0.330

* Significant at the 0.10 level, ** Significant at the 0.05 level, *** Significant at the 0.01 level. Standard errors are in parentheses. The dependent variable is the degree centrality in high school network using (3, 2, 1, 0) tie value set.

Table 4.4. Relationship between Two Centralities in Home Network

VARIABLES	(1) OLS
homeegvc123	6,358***
	(112.8)
Constant	106.2***
	(1.456)
Observations	6,000
R-squared	0.346

* Significant at the 0.10 level, ** Significant at the 0.05 level, *** Significant at the 0.01 level. Standard errors are in parentheses. The dependent variable is the degree centrality in home neighborhood network using (3, 2, 1, 0) tie value set.

Table 4.5 shows analytical results of the high school network employing OLS and Logit models. The results illustrate that coefficients of school degree centrality are always significantly positive, even after I add control variables and the intervening variable "contact_times". There is a positive correlation, therefore, between degree centrality and application decision. Specifically, increasing one standard deviation of high school degree centrality will lead to a 1.4% increase in the application rate, which indicates that a student

who is well connected within his or her high school network is more likely to apply to Baylor University. It suggests that Baylor University has attracted applicants who are clustered by ZIP Codes.

Table 4.6 reveals regression results in home neighborhood networks, in which students form peer groups through their parents, church interactions, childhood friends, their high schools, or in other ways. It can also be found that the coefficients of degree centrality are significantly positive, and the OLS regression results show that increasing one standard deviation of home neighborhood degree centrality will lead to about 1.6% increase in the application rate. Essentially, if a student lives in an area where lots of peers are also identified by Baylor University as potential recruits, he or she will be more likely to apply to Baylor. Therefore, home neighborhood networks are directly relational to student's college choice.

While this paper provides evidence of the importance of peer groups on the college application process, there are several alternative findings which deserve to be mentioned. From our results, it is reasonable to conclude that a student who has more personal contacts with Baylor is significantly more likely to apply. Also, the coefficients of the intervening variable "contact times," which refers to how many times the student connects with Baylor via email, phone, or campus visit, are significantly positive. Thus, Baylor University should take advantage of this finding by increasing interactions with potential recruits in order to raise their probability of applying to Baylor.

Table 4.5. Regression of School Network

VARIABLES	(1) OLS1	(2) OLS2	(3) OLS3	(4) Logit1	(5) Logit2	(6) Logit3
schooldeg123	0.000590***	0.000106***	0.000101***	0.00296***	0.00192***	0.00188***
	(3.92e-05)	(2.02e-05)	(2.02e-05)	(0.000205)	(0.000439)	(0.000441)
gender		-0.0235***	-0.0233***		-0.465***	-0.480***
		(0.00561)	(0.00559)		(0.132)	(0.133)
white		0.0610***	0.0598***		1.680***	1.678***
		(0.00616)	(0.00614)		(0.188)	(0.188)
hispanic		0.223***	0.215***		5.006***	5.007***
		(0.0138)	(0.0138)		(0.384)	(0.385)
asian		0.0981***	0.0957***		2.210***	2.208***
		(0.0136)	(0.0136)		(0.287)	(0.289)
black		0.152***	0.147***		2.733***	2.717***
		(0.0101)	(0.0101)		(0.213)	(0.214)
amerind		0.0923***	0.0920***		2.494***	2.504***
		(0.0209)	(0.0209)		(0.436)	(0.437)
pacisle		0.0175	0.0151			
		(0.0649)	(0.0647)		(omitted)	(omitted)
stype_c		-0.00226	-0.00312		0.187	0.179
		(0.0212)	(0.0212)		(0.692)	(0.690)
stype_public		0.0163	0.0151		0.559	0.535
		(0.0196)	(0.0196)		(0.648)	(0.648)
stype_other		-0.00441	-0.00394		0.231	0.217
		(0.0259)	(0.0258)		(0.792)	(0.790)
hs25_50		0.794***	0.761***		7.004***	7.110***
		(0.0123)	(0.0132)		(0.728)	(0.774)
hs50_90		0.795***	0.762***			
		(0.0128)	(0.0137)		(omitted)	(omitted)
hs90_max		0.791***	0.759***			
		(0.0209)	(0.0214)		(omitted)	(omitted)
total_contacts		0.00931***	-0.00677**		0.323***	-0.293**
		(0.00126)	(0.00276)		(0.0409)	(0.135)
contacttimes			0.0266***			0.672***
			(0.00406)			(0.152)
Constant	0.170***	0.0206	0.00838	-1.540***	-4.873***	-4.913***
	(0.00762)	(0.0220)	(0.0220)	(0.0448)	(0.701)	(0.701)
Observations	6,000	5,943	5,943	6,000	5,336	5,336
R-squared	0.036	0.760	0.762			

* Significant at the 0.10 level, ** Significant at the 0.05 level, *** Significant at the 0.01 level. Standard errors are in parentheses. The dependent variable is a binary variable indicating whether or not the student applied to Baylor University.

Table 4.6. Regression of Home Network

VARIABLES	(1) OLS1	(2) OLS2	(3) OLS3	(4) Logit1	(5) Logit2	(6) Logit3
homedeg123	0.000612***	0.000123***	0.000115***	0.00305***	0.00227***	0.00223***
	(4.05e-05)	(2.09e-05)	(2.09e-05)	(0.000210)	(0.000441)	(0.000443)
gender		-0.0238***	-0.0236***		-0.477***	-0.493***
		(0.00561)	(0.00559)		(0.133)	(0.133)
white		0.0610***	0.0597***		1.688***	1.686***
		(0.00616)	(0.00614)		(0.188)	(0.188)
hispanic		0.222***	0.214***		5.008***	5.009***
		(0.0138)	(0.0138)		(0.385)	(0.386)
asian		0.0970***	0.0946***		2.194***	2.192***
		(0.0136)	(0.0136)		(0.287)	(0.289)
black		0.151***	0.147***		2.724***	2.708***
		(0.0101)	(0.0101)		(0.214)	(0.215)
amerind		0.0918***	0.0916***		2.471***	2.482***
		(0.0209)	(0.0208)		(0.436)	(0.437)
pacisle		0.0193	0.0166			
		(0.0649)	(0.0647)		(omitted)	(omitted)
stype_c		0.00618	0.00478		0.398	0.384
		(0.0213)	(0.0212)		(0.685)	(0.683)
stype_public		0.0242	0.0225		0.763	0.731
		(0.0196)	(0.0196)		(0.640)	(0.640)
stype_other		0.00124	0.00129		0.377	0.355
		(0.0258)	(0.0258)		(0.786)	(0.784)
hs25_50		0.794***	0.761***		7.025***	7.122***
		(0.0123)	(0.0132)		(0.729)	(0.772)
hs50_90		0.795***	0.762***			
		(0.0128)	(0.0137)		(omitted)	(omitted)
hs90_max		0.791***	0.759***			
		(0.0209)	(0.0214)		(omitted)	(omitted)
total_contacts		0.00925***	-0.00686**		0.324***	-0.295**
		(0.00126)	(0.00276)		(0.0408)	(0.136)
contacttimes			0.0266***			0.674***
			(0.00405)			(0.153)
Constant	0.173***	0.0125	0.000562	-1.524***	-5.094***	-5.127***
	(0.00749)	(0.0221)	(0.0221)	(0.0439)	(0.699)	(0.699)
Observations	6,000	5,943	5,943	6,000	5,336	5,336
R-squared	0.037	0.761	0.762			

* Significant at the 0.10 level, ** Significant at the 0.05 level, *** Significant at the 0.01 level. Standard errors are in parentheses. The dependent variable is a binary variable indicating whether or not the student applied to Baylor University.

Robustness Check

Besides the above robustness check in which my regression specification was modified by adding regressors, I also employed another tie value set in building adjacency matrices to examine how consistently my interested coefficient estimates behaved. So instead of the tie value set (3, 2, 1, 0), another set, with different scales (1, 2/3, 1/3, 0) in which 1 refers to the closest relationship while 0 means the weakest connection, was also adopted in the analysis.

Tables 4.7 through 4.8 present regression results using the new tie scale. As we can see, the degree centrality in both networks still has significantly positive effects on the application decision. Similarly, a one standard deviation increase in degree centrality implies about a 1.5% rise in application rates, which means highly connected students with a higher degree centrality are more likely to apply to Baylor University. These robust coefficients consistently help to strengthen the structural validity of my regression results.

Table 4.9 employs another method as a robustness check. The main disadvantage of employing SNA tools in our study is that we lose most of our observations while building relationship matrices. However, the number of students having specific ZIP Codes can be easily calculated throughout all of the 122,967 students in the population who graduated from high school in 2014. Table 4.9 shows multiple regression results using the number of students who have the same five, four, or three digits high school or home ZIP Codes as our interested independent variable. These consistently positive coefficients in regression results further suggest that Baylor University has attracted applicants who are clustered by ZIP Codes.

Table 4.7. Robustness Check of School Network

VARIABLES	(1) OLS1	(2) OLS2	(3) OLS3	(4) Logit1	(5) Logit2	(6) Logit3
schooldeg033	0.00177***	0.000317***	0.000290***	0.00888***	0.00575***	0.00563***
	(0.000118)	(6.07e-05)	(6.07e-05)	(0.000614)	(0.00132)	(0.00132)
gender		-0.0235***	-0.0233***		-0.465***	-0.480***
		(0.00561)	(0.00559)		(0.132)	(0.133)
white		0.0610***	0.0598***		1.680***	1.678***
		(0.00616)	(0.00614)		(0.188)	(0.188)
hispanic		0.223***	0.215***		5.006***	5.007***
		(0.0138)	(0.0138)		(0.384)	(0.385)
asian		0.0981***	0.0957***		2.210***	2.208***
		(0.0136)	(0.0136)		(0.287)	(0.289)
black		0.152***	0.147***		2.733***	2.717***
		(0.0101)	(0.0101)		(0.213)	(0.214)
amerind		0.0923***	0.0920***		2.494***	2.504***
		(0.0209)	(0.0209)		(0.436)	(0.437)
pacisle		0.0175	0.0151			
		(0.0649)	(0.0647)		(omitted)	(omitted)
stype_c		-0.00226	-0.00312		0.187	0.179
		(0.0212)	(0.0212)		(0.692)	(0.690)
stype_public		0.0163	0.0151		0.559	0.535
		(0.0196)	(0.0196)		(0.648)	(0.648)
stype_other		-0.00441	-0.00394		0.231	0.217
		(0.0259)	(0.0258)		(0.792)	(0.790)
hs25_50		0.794***	0.761***		7.004***	7.110***
		(0.0123)	(0.0132)		(0.728)	(0.774)
hs50_90		0.795***	0.762***			
		(0.0128)	(0.0137)		(omitted)	(omitted)
hs90_max		0.791***	0.759***			
		(0.0209)	(0.0214)		(omitted)	(omitted)
total_contacts		0.00931***	-0.00677**		0.323***	-0.293**
		(0.00126)	(0.00276)		(0.0409)	(0.135)
contacttimes			0.0266***			0.672***
			(0.00406)			(0.152)
Constant	0.170***	0.0206	0.00838	-1.540***	-4.873***	-4.913***
	(0.00762)	(0.0220)	(0.0220)	(0.0448)	(0.701)	(0.701)
Observations	6,000	5,943	5,943	6,000	5,336	5,336
R-squared	0.036	0.760	0.762			

* Significant at the 0.10 level, ** Significant at the 0.05 level, *** Significant at the 0.01 level. Standard errors are in parentheses. The dependent variable is a binary variable indicating whether or not the student applied to Baylor University.

Table 4.8. Robustness Check of Home Network

VARIABLES	(1) OLS1	(2) OLS2	(3) OLS3	(4) Logit1	(5) Logit2	(6) Logit3
homedeg033	0.00184***	0.000369***	0.000345***	0.00914***	0.00680***	0.00670***
	(0.000122)	(6.27e-05)	(6.26e-05)	(0.000630)	(0.00132)	(0.00133)
gender		-0.0238***	-0.0236***		-0.477***	-0.493***
		(0.00561)	(0.00559)		(0.133)	(0.133)
white		0.0610***	0.0597***		1.688***	1.686***
		(0.00616)	(0.00614)		(0.188)	(0.188)
hispanic		0.222***	0.214***		5.008***	5.009***
		(0.0138)	(0.0138)		(0.385)	(0.386)
asian		0.0970***	0.0946***		2.194***	2.192***
		(0.0136)	(0.0136)		(0.287)	(0.289)
black		0.151***	0.147***		2.724***	2.708***
		(0.0101)	(0.0101)		(0.214)	(0.215)
amerind		0.0918***	0.0916***		2.471***	2.482***
		(0.0209)	(0.0208)		(0.436)	(0.437)
pacisle		0.0193	0.0166			
		(0.0649)	(0.0647)		(omitted)	(omitted)
stype_c		0.00618	0.00478		0.398	0.384
		(0.0213)	(0.0212)		(0.685)	(0.683)
stype_public		0.0242	0.0225		0.763	0.731
		(0.0196)	(0.0196)		(0.640)	(0.640)
stype_other		0.00124	0.00129		0.377	0.355
		(0.0258)	(0.0258)		(0.786)	(0.784)
hs25_50		0.794***	0.761***		7.025***	7.122***
		(0.0123)	(0.0132)		(0.729)	(0.772)
hs50_90		0.795***	0.762***			
		(0.0128)	(0.0137)		(omitted)	(omitted)
hs90_max		0.791***	0.759***			
		(0.0209)	(0.0214)		(omitted)	(omitted)
total_contacts		0.00925***	-0.00686**		0.324***	-0.295**
		(0.00126)	(0.00276)		(0.0408)	(0.136)
contacttimes			0.0266***			0.674***
			(0.00405)			(0.153)
Constant	0.173***	0.0125	0.000562	-1.524***	-5.094***	-5.127***
	(0.00749)	(0.0221)	(0.0221)	(0.0439)	(0.699)	(0.699)
Observations	6,000	5,943	5,943	6,000	5,336	5,336
R-squared	0.037	0.761	0.762			

* Significant at the 0.10 level, ** Significant at the 0.05 level, *** Significant at the 0.01 level. Standard errors are in parentheses. The dependent variable is a binary variable indicating whether or not the student applied to Baylor University.

Table 4.9. Robustness Check Using the Population

VARIABLES	(1) Same5	(2) Same4	(3) Same3	(4) Same5	(5) Same4	(6) Same3	(7) All
students_in_five_digits_schoolzip	0.000560***						2.80e-05
	(1.10e-05)						(1.98e-05)
students_in_four_digits_schoolzip		0.000245***					7.73e-05***
		(3.94e-06)					(8.97e-06)
students_in_three_digits_schoolzip			3.44e-05***				8.38e-06***
			(5.76e-07)				(1.96e-06)
students_in_five_digits_homezip				0.000721***			0.000140***
				(1.47e-05)			(2.44e-05)
students_in_four_digits_homezip					0.000278***		0.000114***
					(4.38e-06)		(1.03e-05)
students_in_three_digits_homezip						3.39e-05***	1.07e-06
						(5.80e-07)	(2.00e-06)
Constant	0.204***	0.186***	0.194***	0.214***	0.187***	0.200***	0.180***
	(0.00166)	(0.00172)	(0.00165)	(0.00169)	(0.00175)	(0.00167)	(0.00188)
Observations	122,041	122,041	122,041	115,867	119,530	119,530	115,365
R-squared	0.021	0.031	0.028	0.020	0.033	0.028	0.035

* Significant at the 0.10 level, ** Significant at the 0.05 level, *** Significant at the 0.01 level. Standard errors are in parentheses. The dependent variable is a binary variable indicating whether or not the student applied to Baylor University.

CHAPTER FIVE

Conclusions and Discussions

Conclusions

Enrollment is like the lifeblood of colleges and universities. In recent years, admissions recruiters have been facing a lackluster economy, changing demographics, and increased competition among colleges in the US and even around the world. Looking for enough qualified applicants has made the job of admissions recruiters more and more challenging since past performance is no guarantee of future results. This paper examines the role of social networks in college choice, which is important in considering ways to facilitate college recruitment in the future. This study shows that high school networks and home neighborhood networks are correlated to student college choice. Findings reveal that degree centrality has a positive relationship with students' application decisions. This means that highly connected students with a higher degree centrality are more likely to apply to Baylor University. Essentially, if a student lives or studies in an area where lots of peers are also identified by Baylor University as potential recruits, he or she will be more likely to apply to Baylor. As a result, Baylor's admissions recruiters should pay attention to the circle of peers by increasing connections with highly connected students.

Limitations

Essentially, three potential concerns in this paper include sample size, methodology rationality, and data type. These limitations also shape suggestions for future studies.

Although using the small sample of 6,000 students has the advantage of easy computation and exportation via software, it also raises several questions. Big data, which in our case includes 122,967 observations, is holistic and trustworthy. Extracting a small sample from it loses these benefits, as well as other information. And within those 6,000 students, only 1,498 students finally applied to Baylor, so the majority of them did not apply. This leads to reasonable doubt as to whether our sample can reveal common characteristics of Baylor applicants. Though part of my robustness check helps to solve such a problem, our analysis would be more convincing if we were able to employ SNA tools as well as all observations at the same time.

Another concern about our methodology is using the similarity of ZIP Code digits. This method is less convincing than other methods which use exact distance calculations. Also, as far as I know, this methodology is unprecedented in the literature of college choice and social network analysis. As a result, it can be challenging to determine how rational it is.

Finally, our dataset does not have the specialized nature of network data since it is not a standard type in SNA. It is always the case that the time spent analyzing and modeling data is dwarfed by the time spent getting data ready for analyses; there is no difference in this study. However, even after I transformed our original rectangular dataset to sociomatrices, some common network parameters, such as betweenness centrality and closeness centrality, are not workable, which leads to doubt in the comprehensiveness of my analysis.

Recommendations for Future Studies

This study was not able to obtain the exact distance between each student because of serious quota limitations of Google Maps or Bing Maps. Further research can be conducted to gather such distance information in order to build networks more accurately. In addition, researchers in this study did not have enough information to build other types of networks, such as church or student associations. Future research can consider exploring the relationships between other networks and students' application decisions. Also, our conclusions in this study should be tested using the information of applicants from other schools.

Compared with the many studies focusing on peer influence on college choice, there are few studies employing SNA tools and methodologies to examine the network effect on college choice. Future research needs to explore this topic more thoroughly and extensively from the perspective of SNA.

APPENDICES

APPENDIX A

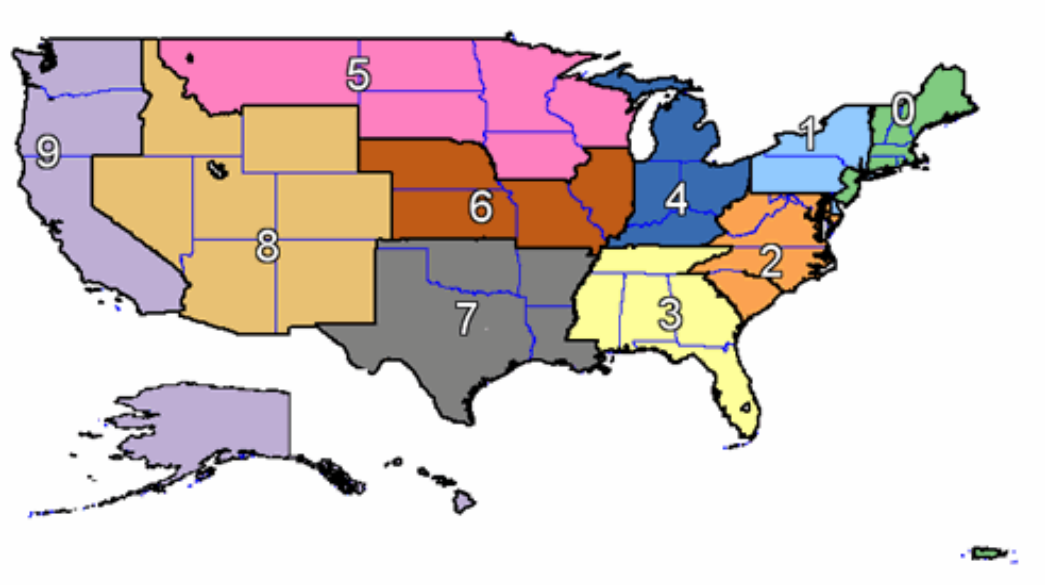

Figure A.1. Geographic Area of ZIP's First Digit.
Note: Picture resource: http://www.zipboundary.com/zipcode_faqs.html

APPENDIX B

Table B.1. Data Dictionary.

Variable	Description
schooldeg	The degree centrality of each node in high school network. Schooldeg123 uses the tie scale of (0, 1, 2, 3). Schooldeg033 uses the tie scale of (0, 1/3, 2/3, 1).
schoolegvc	The eigenvector centrality of each node in high school network. Schoolegvc123 uses the tie scale of (0, 1, 2, 3). Schooldegvc033 uses the tie scale of (0, 1/3, 2/3, 1).
homedeg	The degree centrality of each node in home network. Homedeg123 uses the tie scale of (0, 1, 2, 3). Homedeg033 uses the tie scale of (0, 1/3, 2/3, 1).
homeegvc	The eigenvector centrality of each node in home network. Homeegvc123 uses the tie scale of (0, 1, 2, 3). Homedeg033 uses the tie scale of (0, 1/3, 2/3, 1).
Ethnicity-White Ethnicity-Hispanic Ethnicity-Asian Ethnicity-Black Ethnicity-Amerind Ethnicity-Pacisle	Whether or not the student is white, Hispanic, Asian, black, amerind, or Pacisle (Pacific Islander).
High school type-C High school type-Public High school type-Other	Whether or not the student's high school is Christian non-Public, public or other types.
gender	Male-1, Female-2.
hs_percentile	Percentile of applicant's high school rank. Four levels include min-25, 25-50, 50-90, 90-max.
total_contacts	The total number of contacts between Baylor and the applicant.

contacttimes	How many times the student connected with Baylor via email, phone, or campus visit.
students_in_five_digits_schoolzip	The number of students who have the same five-digit high school ZIP Code
students_in_four_digits_schoolzip	The number of students who have the same four-digit high school ZIP Code
students_in_three_digits_schoolzip	The number of students who have the same three-digit high school ZIP Code
students_in_five_digits_homezip	The number of students who have the same five-digit permanent home ZIP Code
students_in_four_digits_homezip	The number of students who have the same four-digit permanent home ZIP Code
students_in_three_digits_homezip	The number of students who have the same three-digit permanent home ZIP Code

Table B.2. Summary Statistics of Centrality.

Variable	Obs	Mean	Std. Dev.	Min	Max
schooldeg123	6000	135.0193	139.8945	0	523
schoolegvc123	6000	.0031069	.0125316	0	.0635049
homedeg123	6000	125.963	135.3859	0	525
homeegvc123	6000	.0031011	.012533	0	.0667111
schooldeg033	6000	45.00644	46.6315	0	174.3333
schoolegvc033	6000	.0031069	.0125316	0	.0635049
homedeg033	6000	41.98767	45.12865	0	175
homeegvc033	6000	.0031011	.012533	0	.0667111

APPENDIX C

STATA Code – Data Processing

```
# gender
gen      male = (gender=="M")
gen      female = (gender=="F")
gen      gender_missing = (gender == "N")
global   gender male gender_missing

# ethnicity
ren      eth_asian asian
ren      eth_his hispanic
ren      eth_black black
ren      eth_amerind amerind
ren      eth_white white
ren      eth_pacisle pacisle
global   ethnicity asian hispanic black amerind pacisle

foreach x of varlist $ethnicity {
         replace `x' = "1" if `x' == "Y"
         replace `x' = "0" if `x' == ""
         destring `x', replace
         }
gen      ethnicity_missing = (ethnicity == "")

global   ethnicity asian hispanic black amerind pacisle ethnicity_missing
replace ethnicity = "Missing" if ethnicity == ""
encode           ethnicity, gen(eth_dum)
numlabel eth_dum, add

# high school type
replace school_type = "Missing" if school_type == ""
encode school_type, gen(schooltype)
numlabel schooltype, add
gen      stype_c = (schooltype == 3)
gen      stype_missing = (schooltype == 5)
gen      stype_public = (schooltype == 8)
gen      stype_other = (stype_c==0 & stype_missing == 0 & stype_public == 0)
```

```
#high school percentile
sum     high_school_percentile, de
gen     hsmin_25 = high_school_percentile>=r(min) & high_school_percentile<r(p25)
gen     hs25_50  = high_school_percentile>=r(p25) & high_school_percentile<r(p50)
gen     hs50_90  = high_school_percentile>=r(p50) & high_school_percentile<r(p90)
gen     hs90_max = high_school_percentile>=r(p90) & high_school_percentile<=r(max)


#school zip (schoolzip1, schoolzip2, schoolzip3)
replace school_zip = "Missing" if school_zip == ""
encode  school_zip, gen(schoolzip1)
nsplit schoolzip1, digits(4 1) gen(schoolzip2 schoolzipdelete)
nsplit schoolzip2, digits(3 1) gen(schoolzip3 schoolzipdelete2)

#home zip (homezip1, homezip2, homezip3)
replace home_zip = "Missing" if home_zip == ""
encode  home_zip, gen(homezip1)
nsplit homezip1, digits(4 1) gen(homezip2 homezipdelete)
nsplit homezip2, digits(3 1) gen(homezip3 homezipdelete2)
```

APPENDIX D

Matlab Code – Matrix Building Using Two Tie Scales

data type

zip1	zip2	zip3
76706	7670	767
02139	0213	021

```
#(0, 1, 2, 3) tie scale
z1=zeros(length(zip1),length(zip1));
for i=1:length(zip1)
if zip1(i)~=0
y=zip1(i);
v=find(~(zip1-y));
w=zeros(size(zip1));
w(v)= sqrt(3);
q=w*w';
z1=q+z1;
zip1(v)=0;
end
end

z2=zeros(length(zip2),length(zip2));
for i=1:length(zip2)
if zip2(i)~=0
y=zip2(i);
v=find(~(zip2-y));
w=zeros(size(zip2));
w(v)=sqrt(2);
q=w*w';
z2=q+z2;
zip2(v)=0;
end
end

z3=zeros(length(zip3),length(zip3));
for i=1:length(zip3)
if zip3(i)~=0
y=zip3(i);
v=find(~(zip3-y));
w=zeros(size(zip3));
w(v)=sqrt(1);
```

```
q=w*w';
z3=q+z3;
zip3(v)=0;
end
end
matrix=z1+z2+z3;
for i=1:length(zip1)
    matrix(i,i) = 0;
end

filename = 'matrix123.xlsx';
xlswrite(filename,matrix,1)

#(0, 1/3, 2/3, 1) tie scale
z1=zeros(length(zip1),length(zip1));
for i=1:length(zip1)
if zip1(i)~=0
y=zip1(i);
v=find(~(zip1-y));
w=zeros(size(zip1));
w(v)= 1;
q=w*w';
z1=q+z1;
zip1(v)=0;
end
end

z2=zeros(length(zip2),length(zip2));
for i=1:length(zip2)
if zip2(i)~=0
y=zip2(i);
v=find(~(zip2-y));
w=zeros(size(zip2));
w(v)=sqrt(2/3);
q=w*w';
z2=q+z2;
zip2(v)=0;
end
end

z3=zeros(length(zip3),length(zip3));
for i=1:length(zip3)
if zip3(i)~=0
y=zip3(i);
v=find(~(zip3-y));
w=zeros(size(zip3));
```

```
w(v)=sqrt(1/3);
q=w*w';
z3=q+z3;
zip3(v)=0;
end
end

matrix=z1+z2+z3;
for i=1:length(zip1)
    matrix(i,i) = 0;
end

filename = 'matrix033.xlsx';
xlswrite(filename,matrix,1)
```

APPENDIX E

R Code –Visualization and Centrality Calculation

```
setwd(wd)
library(igraph)
library(statnet)
my_data = read.csv(file.choose(),header=TRUE,row.names=1)
B = as.matrix(my_data)
gden(B)
components(B)
gplot(B,gmode="graph",displaylabels=TRUE,edge.lwd=0.8)
gplot(B,gmode="graph",mode="random",vertex.cex=1.5,main="Random layout")
gplot(B,gmode="graph",mode="fruchtermanreingold",vertex.cex=1.5,main="Fruchterma
n-Reingold")
deg = degree(B,gmode="graph")
deg
summary(deg)
cls = closeness(B,gmode="graph")
cls
summary(cls)
bet = betweenness(B,gmode="graph")
bet
summary(bet)
egvc= evcent(B,gmode="graph")
egvc
summary(egvc)
write.table(data.frame(deg), "degree.xls", col.names = TRUE, row.names = FALSE)
write.table(data.frame(egvc), "egvc.xls", col.names = TRUE, row.names = FALSE)
```

BIBLIOGRAPHY

Astin, Alexander W. "What matters in college." (1993).

Beck, Nathaniel. "Is OLS with a binary dependent variable really ok." *Estimating (mostly)* (2011).

Burleson, Douglas A. "Sexual orientation and college choice: Considering campus climate." *About Campus* 14, no. 6 (2010): 9-14.

Carrell, Scott E., Richard L. Fullerton, and James E. West. "Does your cohort matter? Measuring peer effects in college achievement." *Journal of Labor Economics* 27, no. 3 (2009): 439-464.

Davis III, Charles HF, Regina Deil-Amen, Cecilia Rios-Aguilar, and Manuel Sacramento Gonzalez Canche. "Social Media in Higher Education: A literature review and research directions." (2012).

Ding, Weili, and Steven F. Lehrer. "Do peers affect student achievement in China's secondary schools?." *The Review of Economics and Statistics* 89, no. 2 (2007): 300-312.

Dixon, Paul N., and Nancy K. Martin. "Measuring factors that influence college choice." *NASPA Journal* 29, no. 1 (1991): 31-36.

Dooney, Michael. "Social Media & College Admissions: An Analysis of Facebook's Role in College Admissions & Higher Education Marketing." (2014).

Fletcher, Jason, and Marta Tienda. "Race and ethnic differences in college achievement: Does high school attended matter?." *The Annals of the American Academy of Political and Social Science* 627, no. 1 (2010): 144-166.

Furukawa, Derek Takumi. "College choice influences among high-achieving students: an exploratory case study of college freshmen." (2011).

Granovetter, Mark S. "The strength of weak ties." *American journal of sociology* 78, no. 6 (1973): 1360-1380.

Hanneman, Robert A., and Mark Riddle. "Introduction to social network methods." (2005).

Hearn, James C. "The relative roles of academic, ascribed, and socioeconomic characteristics in college destinations." *Sociology of Education* (1984): 22-30.

Hossler, Don, John Braxton, and Georgia Coopersmith. "Understanding student college choice." *Higher education: Handbook of theory and research* 5 (1989): 231-288.

Jackson, Gregory A. "Financial aid and student enrollment." *The Journal of Higher Education* 49, no. 6 (1978): 548-574.

Johnson, Richard G., Norman R. Stewart, and Charles G. Eberly. "Counselor impact on college choice." *The School Counselor* 39, no. 2 (1991): 84-90.

Kapucu, Naim, Farhod Yuldashev, Fatih Demiroz, and Tolga Arslan. "Social network analysis (SNA) applications in evaluating MPA classes." *Journal of Public Affairs Education* (2010): 541-563.

Kealy, Mary Jo, and Mark L. Rockel. "Student perceptions of college quality: The influence of college recruitment policies." *The Journal of Higher Education* 58, no. 6 (1987): 683-703.

Kern, Carolyn W. Kelpe. "College choice influences: Urban high school students respond." *Community College Journal of Research & Practice* 24, no. 6 (2000): 487-494.

Kessler, S. "Social media plays vital role in reconnecting Japan quake victims with loved ones." (2011).

Knoke, David, and Ronald S. Burt. "Prominence." *Applied network analysis* (1983): 195-222.

Maroulis, Spiro, and Louis M. Gomez. "Does “connectedness” matter? Evidence from a social network analysis within a small-school reform." *Teachers College Record* 110, no. 9 (2008): 1901-1929.

Martin, Tait Jeffrey. *Information processing and college choice: an examination of recruitment information on higher education web sites using the heuristic-systematic model*. Florida State University, 2006.

McNeal Jr, Ralph B. "Extracurricular activities and high school dropouts." *Sociology of education* (1995): 62-80.

Paulsen, Michael B. *College Choice: Understanding Student Enrollment Behavior. ASHE-ERIC Higher Education Report No. 6*. ASHE-ERIC Higher Education Reports, The George Washington University, One Dupont Circle, Suite 630, Dept. RC, Washington, DC 20036-1183, 1990.

Ruane, Regina, and Emmanuel F. Koku. "Social network analysis of undergraduate education student interaction in online peer mentoring settings." *Journal of Online Learning and Teaching* 10, no. 4 (2014): 577.

Sacerdote, Bruce. "Peer effects with random assignment: Results for Dartmouth roommates." *The Quarterly journal of economics* 116, no. 2 (2001): 681-704.

Scott, John. *Social network analysis*. Sage, 2012.

Somers, Patricia, Kevin Haines, Barbara Keene, Jon Bauer, Marcia Pfeiffer, Jennifer McCluskey, Jim Settle, and Brad Sparks. "Towards a theory of choice for community college students." *Community College Journal of Research and Practice* 30, no. 1 (2006): 53-67.

Stark, David, and Balázs Vedres. "Social Times of Network Spaces: Network Sequences and Foreign Investment in Hungary 1." *American journal of sociology* 111, no. 5 (2006): 1367-1411.

Summers, Anita A., and Barbara L. Wolfe. "Do schools make a difference?." *The American Economic Review* 67, no. 4 (1977): 639-652.

Tierney, Michael L. "Student college choice sets: Toward an empirical characterization." *Research in Higher Education* 18, no. 3 (1983): 271-284.

Valente, Thomas W., Kathryn Coronges, Cynthia Lakon, and Elizabeth Costenbader. "How correlated are network centrality measures?." *Connections (Toronto, Ont.)* 28, no. 1 (2008): 16.

Wasserman, Stanley, and Katherine Faust. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press, 1994.

Zemsky, Robert, and Penney Oedel. "The structure of college choice." (1983).

Zimmerman, Barry J. "Self-efficacy: An essential motive to learn." *Contemporary educational psychology* 25, no. 1 (2000): 82-91.

Zimmerman, David J. "Peer effects in academic outcomes: Evidence from a natural experiment." *Review of Economics and statistics* 85, no. 1 (2003): 9-23.