Journal of Public Economics 92 (2008) 329–347

JOURNAL OF PUBLIC ECONOMICS

www.elsevier.com/locate/econbase

The old boy (and girl) network: Social network formation on university campuses

Adalbert Mayer, Steven L. Puller *,[1]

Department of Economics, Texas A&M University, United States

Received 2 January 2007; received in revised form 28 August 2007; accepted 4 September 2007
Available online 7 September 2007

Abstract

This paper documents the structure and composition of social networks on university campuses and investigates the processes that lead to their formation. Using administrative data and information from Facebook.com, we document the factors that are the strongest predictors of whether two students are friends. Race is strongly related to social ties, even after controlling for a variety of measures of socioeconomic background, ability, and college activities. We develop a model of the formation of social networks that decomposes the formation of social links into effects based upon the exogenous school environment and effects of endogenous choice arising from preferences for certain characteristics in one's friends. We use student-level data from an actual social network to calibrate the model. We simulate the social network under alternative university policies aimed at reducing social segmentation. We find that changes in the school environment that affect the likelihood that two students interact have only a limited potential to reduce the racial segmentation of the social network.

JEL classification: I20; J15; J62; Z13
Keywords: Social networks; Higher education; Racial segregation

1. Introduction

Universities are important venues for the formation of social networks. For many students, college life is the first experience outside the environment determined by their parents. The resulting social contacts can have far-reaching impacts. They are an important channel of information transmission.[2] Connections between business partners are

* Corresponding author.
E-mail addresses: amayer@econmail.tamu.edu (A. Mayer), puller@econmail.tamu.edu (S.L. Puller).

[1] We are grateful to Larry Malota, Dustin Moscovitz, and Arnie Vedlitz for their time and assistance in acquiring our data. Ahmad Alwaked, Manuel Hernandez, Travis Miller, Lauren Rozanski, and Joseph Wood provided excellent research assistance. We thank two anonymous referees, the Editor Sören Blomquist, Philip Babcock, Julian Betts, Tim Gronberg, Dan McFarland, Kimon Ioannides, John Moroney, Bruce Sacerdote, Adriaan Soetevent, Lori Taylor, Manuelita Ureta, and seminar participants at the 2006 SOLE meetings, the NBER Education Program meeting, IZA/SOLE Transatlantic Meeting, Texas A&M, Rice/University of Houston, the Wegmans Conference at Rochester, University of Augsburg, Western Ontario, the Bush School, Ohio State, University of Texas, University of Montreal, UNC-Greensboro, and University of Calgary for their helpful suggestions. An expanded version of this paper is Texas A&M Department of Economics Working Paper No. 2007-03, which contains robustness tests, empirical analysis of additional network features, and a more detailed description of the data.

[2] For surveys of the social interactions literature, see Ioannides and Loury (2004), Manski (2000), and Soetevent (2004).

0047-2727/$ - see front matter
doi:10.1016/j.jpubeco.2007.09.001

formed using knowledge from prior social interaction.[3] Employers and employees frequently use social contacts to obtain information about each other, which can have important impacts on labor markets.[4]

Interaction between members of different social groups in college is sometimes viewed as a policy goal in itself. It is argued that interaction between students from different backgrounds and walks of life provides a better learning environment to prepare students for an increasingly diverse workforce and society (see Bok and Bowen, 1998). Universities have made concerted efforts over the last several decades to create diverse campuses that bring together students of different races and from diverse socioeconomic backgrounds. In a recent ruling (Grutter v. Bollinger, 2003) the United States Supreme Court upheld the right of universities to use race as an admissions criterion to increase diversity. One argument against the current practice is the disconnect between the admissions policies and the actual campus experience.[5]

Indeed, it is not known whether a diverse university population leads to diverse interaction among students. Anecdotal evidence suggests that students form cliques based upon race or social background — a casual walk through a university campus or a visit to the dorm cafeteria will illustrate social segmentation. There are few large scale empirical studies that quantify this pattern. A notable exception is Marmaros and Sacerdote (2006), who use data on email communication between Dartmouth students and find that race and residential proximity are important determinants of social interaction.[6]

In this paper, we use a large new dataset from 10 public and private universities to describe social networks in college. This paper empirically models the process of social network formation and applies the model to address public policy concerns regarding social segmentation. We make two contributions to the literature. First, we use student-level data to document the structure of these networks and to measure segmentation of social ties by race, socioeconomic background, and ability. Second, we develop a model of network formation that yields a network with many of the commonly observed characteristics of social networks. We calibrate the model to our data and perform counterfactual experiments of university policies that promote student diversity.

Our data are from Facebook.com, a student social networking website for each university. Students use this website to share information and stay in contact with each other. One feature of the Facebook identifies friendships between students and we exploit this information to measure students' social connections on campus.

We find that social networks differ substantially from the network that would arise from the random selection of friends. The structure of these networks exhibits the classic characteristics of social networks (see Jackson, 2006); social networks on campus are cliquish, the distribution of number of social connections is right skewed, and agents with many ties tend to be connected with other agents with many ties.

At all 10 universities, similar characteristics of two students make the formation of a friendship more likely. Despite the fact these schools are very different in size and type, we find similar overall patterns in social segmentation. Two students are moderately more likely to form a friendship if they share the same major or political orientation or belong to the same cohort. However, friendships are much more likely to be formed within the same race for minorities.

We match the social network data to student-level administrative data for one of the universities—Texas A&M. This allows us to use additional information on parental education and income, student SAT scores, high school, college GPA, dorm, and activities, such as athletics and fraternity/sorority membership.

Using this rich dataset on Texas A&M students, we investigate the determinants of friendship formation. First, we explore in a reduced-form setting the demographic and socioeconomic factors that are good predictors of two students becoming friends. Second, we use the reduced-form results to develop a model of friendship formation that we calibrate to our data.

[3] For example, Cohen et al. (2007) study the impact of shared education networks between mutual fund managers and corporate board members on investment choices.

[4] Montgomery (1991) reviews several studies about the importance of social connections in the labor market and concludes that: "While the frequency of alternative job-finding methods varies somewhat by sex and occupation, the following generalization seems fair: approximately 50% of all workers currently employed found their jobs through friends and relatives". Pellizarri (2004) documents this observation for various countries. Calvo-Armengol and Jackson (2004, 2007) present consequences of this phenomenon.

[5] In his dissenting opinion Justice Scalia states: "Still other suits may challenge the bona fides of the institution's expressed commitment to the educational benefits of diversity that immunize the discriminatory scheme in *Grutter*. (Tempting targets, one would suppose, will be those universities that talk the talk of multiculturalism and racial diversity in the courts but walk the walk of tribalism and racial segregation on their campuses…)".

[6] For analyses of secondary school social interaction and connectedness, see Joyner and Kao (2000), Moody (2001), Quillian and Campbell (2003), Weinberg (2005), Fryer and Torelli (2006), and Babcock (2006).

We estimate a linear probability model of any two students being friends. Relative to the baseline rate that any two students chosen at random are friends, students living in the same dorm are 13 times more likely to be friends, two Black students are 17 times more likely to be friends, two Asian students are 5 times more likely to be friends, and two Hispanic students are about twice as likely to be friends. Socioeconomic background and academic achievement affect the probability of a friendship formation to a smaller but statistically significant degree.

Even though observable characteristics such as race clearly play a role in friendship formation, they have very little explanatory power for the formation of a friendship *between two specific students*. However, common friends are a good predictor for the existence of a friendship between two students — students i and j are much more likely to be friends if each is friends with student k. Moreover, when we control for the number of common friends, the importance of other characteristics such as race changes. Therefore, a linear probability model leads to biased predictions of the effects of changes in the school environment. A better model must incorporate how common friends can affect the formation of a friendship.

Thus, we build a model of friendship formation. This allows us to simultaneously study the environment that determines whether students meet, tastes that determine the formation of friendships conditional on meeting, and the influence of interconnections through friends of friends. Like Jackson and Rogers (2007), we allow links to form with any other student ("random attachment") and links to form through friends of existing friends ("preferential attachment" or "search")[7]. In addition, we model heterogeneity in the environment and preferences of agents.

Our model starts with a network in which no students are connected. Then, two individuals meet with a probability that is determined by their school environment (e.g. dorm assignment or cohort). Conditional upon meeting, the students choose whether or not to form a friendship based upon tastes for observable characteristics. Finally, students meet friends of their friends and again choose whether to form a friendship based upon preferences. We calibrate the model by simulating a network that resembles the observed social network.

We simulate modified versions of the model to generate policy counterfactuals. We assess the effectiveness of policies that try to decrease socialization *within* subgroups and increase socialization *across* subgroups. Our experiments suggest that there is very little potential to increase the social ties between different groups by changing the environment that leads to contact between students. Segmentation by race or background appears to be mainly driven by preferences to form friendships conditional on meeting, and less by differences in the probability of meeting. We also simulate the effects of an increase in the number of minority students. The model predicts more segmentation of the minority group in question, however, the fraction of minority friends of other races increases.

Much of the existing literature that studies the effects of one's academic peers exploits unique settings with random assignment, such as the assignment of roommates (e.g. Sacerdote, 2001; Marmaros and Sacerdote, 2006). Unfortunately, settings with random assignment of peers are rare for the researcher. Most of a typical student's peers are not assigned randomly but are chosen through a process of network formation. We adapt methods from the social networks literature to model the process by which a student forms her peers. This literature has developed models of network formation with the goal of specifying parsimonious models that explain certain features of observed social networks such as clustering and small-world effects. Our goal is somewhat different. We seek to understand how the process of network formation contributes to the segmentation of heterogeneous individuals. Existing social network models do not incorporate heterogeneity in agents, and therefore do not generate the observed segmentation. We add heterogeneity in preferences and environment to evaluate policies aimed at reducing social network segmentation.

In Section 2 we describe our data, document the structure of the networks, and analyze associations between individual characteristics and friendship formation. In Section 3, we present a model of network formation and use it to simulate counterfactual networks in Section 4. In Section 5 we show that a student's academic and social outcomes are related to her friends. Section 6 concludes.

2. Description of social networks

2.1. Data from Facebook.com

Our data on social networks are from the Facebook.com website in early 2005. The online student directory Facebook.com was conceived by undergraduate students at Harvard in February 2004. In spring 2004, the Facebook

[7] A number of mechanical stochastic processes of network formation have been proposed. These models are able to explain various features of social networks. Contributions can be found in the computer science, physics and economics literature (See Newman (2003)).

expanded beyond Harvard to other Ivy League schools and by fall 2004 Facebook.com had added websites for several hundred colleges and universities around the country. To participate on the Facebook, students must sign up using an official university email address, ensuring that they are members of the campus community.[8] Students set up a profile page that includes a picture, name, gender, high school, major, current classes, political orientation, music tastes, hobbies and other interests. Students use the website to share information and stay in contact with each other.

The students' profiles also contain a list of 'friends'. A Facebook friendship is formed if student A sends a friendship request via the website to student B and student B accepts A's friendship invitation. Student A appears as a friend on B's Facebook profile and vice versa. We use these friend connections as a proxy for a student's social network.

2.1.1. What do 'Facebook-friendships' measure?

Facebook friendships are very likely to measure interaction on campus. In informal surveys, students describe their Facebook friends as acquaintances made at school or social activities. We also can provide slightly more formal evidence. After we collected our data, Facebook added a feature that allows students to self-report how they met each of their friends. Using a sample of this information for Texas A&M, we found that the main channels of meeting friends were being co-members of a school organization (26%), meeting through another friend (16%), attending the same high school (14%), and taking a course together (12%). Very few friendships are merely online interactions (0.4%). We believe that Facebook friends are likely to include not only close friends but also the "weak ties" that Granovetter (1973) describes as being important for information transmission. Section 5 provides further evidence that a student's friends, as measured by Facebook, are associated with educational outcomes.

2.1.2. Universities in our sample

We have a snapshot of data from Facebook websites at 10 Texas universities on January 17, 2005.[9] At these universities, Whites comprise a strong majority of the student population. The largest minority group tends to be Hispanics followed by Asians and Blacks. Facebook does not ask students to report their race. Therefore, we classify the pictures on the Facebook profiles by race. The race categories used in this classification are: White/Hispanic, Black and Asian[10].

Table 1 shows characteristics of students at each university. The schools are ordered by the date at which the Facebook was established on campus. A large fraction of students are registered on the Facebook website at the time our sample was drawn. 80% of undergraduates at Rice, 40% at University of Texas and 44% at Texas A&M are included in our Facebook sample. Altogether, our Facebook sample contains 38,923 undergraduate students.

In Section 2.2 we analyze the Facebook networks at all 10 Universities. In Section 2.3 we match the Facebook data to additional student-level data from administrative records at Texas A&M. This allows us to look at the predictors of friendships in more detail. We also use the additional information to address issues of selection into Facebook.

2.2. Social networks at 10 Texas universities

First, we document the characteristics of the networks at all 10 universities. The Facebook networks exhibit characteristics common to social networks and are strongly segmented by race, cohort, major, and political orientation.

2.2.1. Network structure

A vast literature in sociology, mathematics and computer science provides an array of different tools to characterize networks. In order to present some of these measures, we need to introduce some notation.[11] We consider a campus with n students, or in the terminology of network analysis, a network with n nodes. Students i and j can be friends with

[8] Facebook has changed features over time, so some current features were not available when our data were collected.

[9] The administrators of Facebook.com provided us with data on all student profiles at Rice, University of Texas, Texas A&M, Baylor, Texas Tech, Texas Christian University, Southern Methodist University, University of North Texas, UT-Arlington and Texas State University. We thank Dustin Moscovitz for his assistance in providing us with the data.

[10] Each picture was evaluated by two undergraduate research assistants. We only include students in our analysis if both research assistants' race evaluation coincided. In the working paper we show that the race evaluation by the RAs coincide in the vast majority of cases with official race classification at Texas A&M.

[11] The presentation here is based on Jackson (2006). For other ways to characterize networks, see Newman (2003) and Wasserman and Faust (1994).

Table 1
Network composition and characteristics

	Rice	U Texas	Texas A&M	SMU	Baylor	Texas Tech	Texas Christian	U North Texas	UT-Arlington	Texas State
Facebook uptake rate:	80%	40%	44%	57%	61%	31%	52%	18%	8%	22%
Composition										
Number of students	1300	8467	9299	2223	4295	4648	2342	2607	820	2922
Fraction female	0.50	0.56	0.55	0.59	0.60	0.53	0.64	0.57	0.47	0.58
Fraction White or Hispanics	0.82	0.85	0.96	0.94	0.91	0.97	0.95	0.92	0.82	0.96
Fraction Black	0.05	0.02	0.02	0.04	0.06	0.02	0.04	0.06	0.12	0.03
Fraction Asian	0.13	0.13	0.02	0.03	0.03	0.01	0.01	0.02	0.06	0.01
Fraction liberal	0.32	0.23	0.06	0.11	0.08	0.08	0.12	0.18	0.16	0.14
Fraction conservative	0.15	0.23	0.54	0.42	0.47	0.50	0.43	0.27	0.27	0.31
Characteristics										
Average number friends	50.8	39.5	41.1	62.9	59.8	40.5	49.8	23.8	17.2	25.6
Variance of number of friends	31.9	36.5	38.4	48.3	50.8	35.6	36.0	23.9	17.7	23.8
Skewness of number of friends	1.06	2.01	2.06	1.75	1.74	1.50	1.11	2.28	1.52	1.69
Cluster coefficient	0.24	0.20	0.17	0.23	0.19	0.21	0.23	0.21	0.27	0.23
Degree correlation	0.22	0.57	0.57	0.49	0.58	0.57	0.54	0.35	0.53	0.55

Note: The uptake rate is the number of undergraduates on Facebook divided by the total undergraduate enrollment. Composition and characteristics are reported for students for whom we could identify race based upon the picture. Degree, degree correlation, and cluster coefficient are defined in Section 2.2.

each other, in which case the nodes i and j are linked or connected. This relationship is symmetric; if student i is a friend of student j, then student j is also a friend of student i. The friendships between students are recorded in the symmetric $n*n$ matrix $\boldsymbol{g}$. If student i and student j are friends, $g(i, j)=1$ and $g(j, i)=1$. Otherwise, the elements of $\boldsymbol{g}$ are equal to zero.

One measure of the cliquishness of a network is the *cluster coefficient*. It captures the fraction of the friends of a given individual who are friends with each other. The literature considers different ways of calculating this measure. We follow Jackson and Rogers (2007) and define the *cluster coefficient* of the network as:

$$C = \frac{\sum_{i;j\neq i, k\neq j,i} g_{ij}g_{jk}g_{ik}}{\sum_{i;j\neq i, k\neq j,i} g_{ij}g_{jk}}$$

According to Jackson (2006) and Newman (2003), social networks are characterized by a number of common characteristics. The degree distribution (the distribution of the number of friends) is right skewed and has fat tails. Social networks tend to be cliquish and exhibit a cluster coefficient that cannot be explained by random formation of links.[12] Social networks exhibit positive degree correlations—nodes with many (few) links are connected to other nodes with many (few) links.

The lower half of Table 1 shows characteristics of the Facebook networks at the 10 universities. The standard features of social networks are exhibited. The average number of friends ranges from 17.2 at the UT-Arlington to 62.9 at SMU. This is partially explained by the date that Facebook started on each campus. The variance of the number of friends is closely associated with the mean; it ranges from 17.7 at UT-Arlington to 50.8 at Baylor. The number of friends is clearly right-skewed at all 10 universities.

[12] Newman (2003) and Jackson (2006) report cluster coefficients ranging from .09 to .45 for co-authorship networks in different academic disciplines; Goyal et al. (2006) report cluster coefficients from .16 to .20 among co-authors in economics.

Table 2
Segmentation by race, major, cohort and political orientation

	Rice	U Texas	Texas A&M	SMU	Baylor	Texas Tech	Texas Christian	U North Texas	UT-Arlington	Texas State
Segmentation by race										
Fraction of students White/Hispanics	0.82	0.85	0.96	0.94	0.91	0.97	0.95	0.92	0.82	0.96
Fraction friends of White/Hispanics who are White/Hispanics	0.85	0.93	0.97	0.96	0.96	0.98	0.97	0.94	0.92	0.97
Fraction of students Asian	0.13	0.13	0.02	0.03	0.03	0.01	0.01	0.02	0.06	0.01
Fraction Friends of Asians who are Asian	0.30	0.58	0.16	0.22	0.25	0.07	0.05	0.10	0.23	0.02
Fraction of students Black	0.05	0.02	0.02	0.04	0.06	0.02	0.04	0.06	0.12	0.03
Fraction friends of Blacks who are Black	0.25	0.38	0.27	0.32	0.47	0.17	0.25	0.33	0.58	0.18
Pair of:	Relative probability of friendship									
White/Hispanics and White/Hispanics	1.03	1.12	1.01	1.05	1.10	1.02	1.03	1.04	1.14	1.01
White/Hispanics and Asian	0.79	0.42	0.74	0.61	0.43	0.52	0.55	0.77	0.32	0.84
White/Hispanics and Black	0.87	0.56	0.77	0.53	0.41	0.70	0.65	0.66	0.55	0.75
Asian and Asian	2.41	4.13	7.42	6.24	4.23	3.85	2.45	3.58	1.59	1.78
Asian and Black	0.92	0.54	1.01	0.86	0.52	0.80	0.77	0.69	0.36	1.00
Black and Black	5.12	13.13	16.54	6.92	5.99	7.35	5.59	5.03	5.71	6.33
Segmentation by major										
Fraction of friends in same major if friendships were formed randomly	0.04	0.02	0.02	0.01	0.02	0.03	0.01	0.01	0.05	0.01
Actual fraction of friends in same major	0.08	0.08	0.07	0.08	0.06	0.06	0.07	0.08	0.10	0.08
Segmentation by cohort										
Pair of:	Relative probability of friendship									
Freshman and freshman	2.14	2.24	2.10	2.10	2.10	2.01	1.95	1.85	1.72	2.07
Freshman and sophomore	0.64	0.74	0.72	0.64	0.60	0.82	0.74	0.84	1.00	0.79
Freshman and junior	0.46	0.40	0.45	0.38	0.33	0.52	0.45	0.62	0.73	0.46
Freshman and senior	0.35	0.25	0.31	0.20	0.18	0.43	0.25	0.58	0.61	0.31
Sophomore and sophomore	2.18	2.28	2.04	2.42	2.62	1.80	2.19	1.74	1.29	2.01
Junior and junior	2.17	2.13	2.14	2.21	2.29	1.46	2.17	1.55	1.27	1.77
Senior and senior	1.80	2.05	2.43	2.08	2.06	1.71	1.92	2.38	1.95	1.93
Segmentation by political orientation										
Pair of:	Relative probability of friendship									
Liberal and liberal	1.22	1.06	1.28	1.00	1.13	1.07	1.09	1.18	1.24	1.05
Liberal and conservative	0.86	0.75	0.69	0.66	0.59	0.70	0.85	0.76	0.79	0.81
Conservative and conservative	1.35	2.17	1.28	1.36	1.41	1.44	1.30	1.45	1.84	1.53

Note: This table includes undergraduates in our Facebook.com sample for whom we could identify race based upon the picture. Students were classified as either White/Hispanics, Black, Asian, or Don't Know, as described in Section 2.2. The fraction of pairs of students of race X and Y who are friends is the fraction of all possible pairs of students of race X and Y who report being friends (reported in percentage points). The relative probability of friendship is defined in Section 2.2.

All 10 networks are clustered. The cluster coefficient ranges from 0.17 at Texas A&M to 0.27 at UT-Arlington. Larger networks tend to have a smaller cluster coefficient[13]. The degree correlation is always positive—it ranges from .22 at Rice to .58 at Baylor.

2.2.2. Segmentation of the social networks

Table 2 shows that the friendship networks at the 10 Texas universities are segmented by race, major, cohort, and political orientation. A variety of definitions and measures of segmentation, or segregation, have been proposed in the literature (see Echenique and Fryer (2007) and Newman (2003)). We compare the probability that two members of a

[13] The cluster coefficient is directly related to the network density (the probability that any two nodes are connected). All else equal, denser networks have higher cluster coefficients. In a completely random network, the cluster coefficient is given by the probability that any two students are friends.

Table 3
Factors predicting the probability that two students are friends

Dependent variable = 1 if students i and j are friends and = 0 otherwise
Mean of dependent variable (baseline rate): 0.0034
Relationship between student i and j

	Race	High school, age	Family	Dorm, academic	Ability	Activities	All	Common friends?
Constant	0.0036 **	0.0039 **	0.0023 **	0.0028 **	0.0045 **	0.0032 **	0.0032 **	−0.0005 **
Both Black	0.0551 **						0.0537 **	0.0153 **
Both Asian	0.0122 **						0.0121 **	0.0074 **
Both Hispanics	0.0017 **						0.0021 **	0.0016 **
Both Native American	−0.0036 **						−0.0038 **	−0.0008
White–Hispanics	−0.0010 **						−0.0003 **	0.0001
White–Asian	−0.0012 **						−0.0007 **	0.0000
White–Black	−0.0012 **						−0.0006 **	−0.0004
White–Native American	−0.0010						−0.0008	0.0005
Hispanic–Asian	−0.0009 **						0.0000	0.0003
Hispanic–Black	0.0000						0.0005	0.0000
Hispanic–Native American	−0.0012 *						−0.0005	0.0009
Asian–Black	−0.0002						0.0005	−0.0002
Asian–Native American	−0.0020 **						−0.0014 **	0.0003
Black–Native American	−0.0018						−0.0015	0.0001
Same high school		0.1864 **					0.1859 **	0.1379 **
Same year in college		0.0010 **					0.0011 **	0.0012 **
Same gender		0.0006 **					0.0000	−0.0005 **
Difference b/t years in college (years)		−0.0013 **					−0.0011 **	0.0001 **
Both from high income households (>$60 k)			0.0005 **				0.0002	−0.0003 **
Both from low income households (<$60 k)			0.0003 **				0.0003 **	0.0003 **
2 college parents–2 college parents			0.0013 **				0.0009 **	−0.0013 **
2 college parents–1 college parent			0.0004 **				0.0003	−0.0008 **
1 college parent–1 college parent			0.0002				0.0001	−0.0004 **
2 college parents–0 college parents			−0.0001				0.0000	−0.0006 **
1 college parent–0 college parents			−0.0001				−0.0001	−0.0003 **
Students both liberal			0.0025 **				0.0021 **	0.0017 **
Students both conservative			0.0023 **				0.0019 **	−0.0012 **
Students one liberal one conservative			−0.0002				−0.0001	−0.0003
Same dorm				0.0426 **			0.0406 **	0.0214 **
Same major				0.0038 **			0.0030 **	0.0024 **
Same college on campus				0.0018 **			0.0016 **	0.0004 **
Difference in SAT scores (absolute points in 100 s)					−0.0004 **		−0.0003 **	0.0000
Difference in GPA quintile (0–4 absolute quintiles)					−0.0003 **		−0.0002 **	−0.0001 **
Both are athletes						0.0646 **	0.0633 **	0.0110 **
Both in corps of cadets						0.0531 **	0.0421 **	0.0218 **
Both are Greek						0.0189 **	0.0183 **	−0.0082 **
One is Greek						−0.0003 **	−0.0004 **	−0.0022 **
One is athlete						−0.0003	−0.0003	−0.0015 **
One in corps of cadets						−0.0005	−0.0006	−0.0005
Number of common friends								0.0299 **
R^2	0.0006	0.0293	0.0004	0.0033	0.0001	0.0032	0.0362	0.2457

Obervations are all pairwise combinations of students in Texas A&M Facebook with complete data on covariates ($29{,}787{,}621 = N*(N-1)/2$ observations where $N=7719$). Linear probability model estimated via least squares. Bootstrap confidence intervals are constructed by sampling with replacement over individual students to obtain 7719 students and forming all pairwise combinations of those students as the bootstrap sample. We construct 200 bootstrap samples. Table only reports coefficient estimates and significance levels to conserve space but confidence intervals are available upon request. Excluded category for race is White–White and for political orientation is no reported orientation.

* Significant at 5%.

** Significant at 1%.

subgroup are friends, to the probability that two random students are friends. This measure of relative segmentation is independent of the size of the two different groups. The relative probability of friendship of Blacks, for example, is given by:

$$\text{Relative Probability of Friendship (black and black)} = \frac{\dfrac{\text{Number of pairs of blacks who are friends}}{\text{Total number of pairs of blacks}}}{\dfrac{\text{Number of pairs of any students who are friends}}{\text{Total number of any pairs}}}.$$

Table 2 shows that students of the same race are more likely to form a friendship than students of different races. Most students are White/Hispanic and the probability that two White/Hispanic students form a friendship is similar to friendship formation of any two random students (unity). Two Asian students are 1.59 (at UT-Arlington) to 7.42 (at A&M) times more likely to be friends than any two random students. For pairs of Blacks, this ratio ranges from 5 (at U of North Texas) to 16.5 (at A&M).[14] The relative probability of friendship is smaller than one for cross-race pairs.

The actual social environment of an individual is determined by the likelihood of forming a friendship with a particular race *and* by the racial composition of the student body. The fraction of Black friends of a Black student depends on their relative probability of friendship formation and the share of Blacks in the entire student population:

$$\begin{aligned}&\text{Fraction black friends of black student}\\ &\quad = \text{Relative Probability of Friendship (black and black)} * \text{(share of blacks in population)}.\end{aligned}$$

The top part of Table 2 documents the absolute segmentation. If friendships were formed randomly, the distribution of characteristics among the friends of any subset of students should equal the distribution in the population. At all universities and for all races, students have a higher fraction of friends from their own race than implied by random assignment. For example, 13% of the students from Rice are Asian, but 30% of the friends of Asian students are Asian. 25% of the friends of Blacks at Rice are Black while Blacks comprise only 5% of the student population. While students have disproportionately many friends of the same race, it is also true that students mix across races. Students at more diverse universities have more diverse social networks. For example, White/Hispanic students at institutions with a large share of minorities tend to have more minority friends than Whites/Hispanics at more homogeneous institutions. In general, minorities tend to have more diverse social networks.

Table 2 also documents segmentation by major, cohort, and political orientation. Students have at least twice as many friends from the same major than random friend assignment would generate. Two students in the same cohort are about twice as likely to be friends as two random students. At all schools, self-reported conservatives have disproportionately many conservative friends and liberals disproportionately many liberal friends, however this segmentation is weaker than the race segmentation for minorities.

2.3. Friendship formation at Texas A&M

The 10 Facebook networks described in Section 2.2 are all segmented by race, major, cohort, and political orientation, and they all exhibit standard features of social networks.

From now on we focus on one of these networks, Texas A&M. For this university we have additional information about the students' characteristics. We match data from the Facebook to administrative data from the Texas A&M registrar's office. The administrative data include the academic record of the students (i.e. major, grade point average), race, dorm, membership in sororities and fraternities, and information about parental background, SAT scores and high school. We use administrative data on race rather than the visual race categorization used above. This allows us to distinguish Hispanics, who are the largest minority at Texas A&M.

In order to evaluate sample selection, we also obtained summary statistics of these variables for students not in the Facebook. We compare summary statistics for students in our Facebook sample to the overall student population. The two samples are very similar along most dimensions. Sample means are very similar for GPA (2.95 in Facebook vs.

[14] The segmentation by race is more pronounced for smaller minorities and at bigger institutions. Possible explanations are that smaller minorities stick together; and that a larger absolute number of students facilitates segmentation, as the number of minority students with specific interests increases. Future work could explore these conjectures.

2.93 in the overall population), SAT (1168 vs. 1152), high school percentile (87 vs. 86), and athletic participation (2.5% in each). The Facebook tends to be slightly more popular among female students (55% vs. 51%) and among younger students. The latter feature explains the higher fraction of students living in a dorm in the Facebook sample (41% vs. 34%). Members of fraternities/sororities (Greek) are overrepresented in the Facebook (14% vs. 12%). Two minority groups are slightly underrepresented—Blacks (2.3% vs. 2.9%) and Hispanics (11.4% vs. 12.0%). Students in the Facebook are slightly more likely to have college educated parents (by 3 percentage points) and to come from a household with income over \$80 k (by 5 percentage points).[15]

We construct a sample that contains the 7719 students in the Texas A&M Facebook network for whom we have complete data on race, demographics, family background, SAT scores, GPA and college activities.

We consider all pairs of students (i.e. $\frac{N(N-1)}{2}$ possible friendship pairs) and quantify the relationship between their characteristics and the formation of friendships. We estimate a linear probability model of the form:

$$\mathrm{Friends}_{ij} = f(X_i, X_j, \varepsilon_{ij}; \beta) \qquad \text{for all } i \neq j$$

where $\mathrm{Friends}_{ij}$ is an indicator of whether two students are Facebook friends and X_i, X_j are characteristics of the two students. We do not view this evidence as causal but merely as an analysis of the factors that are good predictors of friendship. The results are shown in Table 3. When we condition on none of the students' characteristics, the probability that any two students are friends is 0.34%. Such a small baseline rate is not surprising for a large university.

In the first column, we analyze the extent to which the race of students i and j serve as predictors of friendship. As seen above, students of the same race are more likely to form friendships. Both students being African American and both being Asian significantly increases the probability of being friends. Two students who are Black are 17 times more likely to be friends than two students chosen at random (i.e. (0.0551 + 0.0036)/0.0034). Two Asian students are 5 times more likely to be friends. The probabilities that pairs consisting of a White and a minority student form a friendship are around .25%, lower than the baseline .34%.

Subsequent regressions in columns 2–6 test for associations between friendship and various other factors: cohort, high school, parental characteristics, institutional characteristics, ability and campus activities. Having attended the same high school increases the likelihood of being friends at college. Students in the same year of school are more likely to be friends and larger differences in years reduce the likelihood of friendships. Students from families with similar income levels are more likely to be friends, as are students if each has at least one parent with a college education. Living in the same dorm leads to about a 13 fold increase in the probability of being friends relative to two randomly chosen students. Other institutional factors, such as being in the same major or the same college, also increase the likelihood of friendship, however these effects are an order of magnitude smaller than living in the same dorm. Two students are slightly less likely to be friends relative to the baseline if they have SAT scores or college GPAs that differ substantially. Finally, campus activities affect the probability of being friends. Relative to the case of neither student participating in a particular activity, students are more likely to be friends if both participate but less likely if one participates and one does not.

In the seventh column, we include all sets of characteristics as predictors. Many of the coefficients in this model are very similar to their counterpart in the model with fewer covariates.

In particular, the coefficients for race are robust to controlling for demographics, ability, dorm, major and activities. The fact that adding covariates does not significantly change the race coefficients suggests that the observed social network segmentation by race does not merely reflect different institutional channels of meeting such as major, athletics or dorm. Rather, it suggests that there are tastes for characteristics that are correlated with race that affect the probability of becoming friends. We incorporate these insights into our model below.

These results suggest that the strongest predictors of friendship are sharing the same high school, same dorm, same race for minority students, same campus organizations, same major/college, same political orientation, being from the same cohort, and to a lesser extent sharing similar parental background characteristics.[16]

Note that in each of the models discussed above, the R^2 is low. This is not surprising—there are many unobserved characteristics, tastes, and coincidences that determine friendship formation. We illustrate the importance of one of

[15] For a full comparison of all sample characteristics, see our working paper.

[16] These results are consistent with the findings of Marmaros and Sacerdote (2006). They report that physical distance of residence and cohort are two important institutional factors that determine the interaction between students. Ward (2004) also studies the effect of distance on interaction.

these additional factors—having common friends. The last column of Table 3 shows estimates of the linear probability model when we include the number of common friends as an additional regressor. Common friends are a good predictor for the existence of a friendship—the R^2 increases from below .04 to almost .25. Moreover, conditioning on the number of common friends changes the importance of other characteristics, such as common race. The coefficient for both Black drops by two-thirds and the coefficients for both Asian and both Hispanic drop by one-third and one-quarter, respectively. It is difficult to interpret the meaning of these changes, as the formation of friendships and the determination of friends of friends are the outcomes of the same process.

This suggests that if we want to assess effects of changes in any of the variables determining friendship formation, it is not sufficient to use the results of the linear probability model. Endogenous effects through friends of friends may magnify effects of a changed environment. The probability that student i and j meet is a function of whether they are both friends with student k, so characteristics of student k also affect the probability that i and j are friends. Therefore, we proceed by building a model of friendship formation.

3. Model of social network formation

We seek to understand the process of network formation and quantify the importance of different determinants of the network, while taking endogenous network effects into account. A model of social network formation makes it possible to evaluate policies that alter social interactions in college. For example, such a model could evaluate the effect of increasing the number of ethnic minorities, admitting students from different parts of the ability distribution, or changing freshman dorm assignments.

Our model combines a stochastic meeting process and choices by individuals based on their preferences. Like in Jackson and Rogers (2007), the meeting process consists of random encounters and introduction to friends of friends. We add heterogeneous agents and a simple preference structure. We calibrate our model to fit data on an actual network and conduct counterfactual experiments.

3.1. Mechanics of the model

The model starts with a completely unconnected network – no friendships have been formed and all elements of the friendship matrix $\boldsymbol{g}$ are equal to zero. A friendship between students i and j is the outcome of two events: (1) two students meet with some probability, and (2) conditional upon meeting, students choose whether or not to form a friendship. Students i and j meet each other with a probability $p_{ij}(Z_i,Z_j)$, which is a function of observable features of each student's institutional environment $Z_{i/j}$ (e.g. living in the same dorm or being part of the same cohort). In addition, students meet other students through their existing friends.

After two students meet, they decide whether they like each other.[17] This decision depends on one another's characteristics, some of which are observable ($X_{i/j}$) and some of which are unobservable ($u_{i/j}$) to the analyst. Denote U_{ij} ($X_i,X_j,u_i,u_j;\beta$) the utility student i derives from being friends with student j, and c_i the marginal cost of friendship to student i (e.g. the time cost of a friendship). Because friendship is mutual, we model friendship as:

$$\begin{aligned} g(i,j) &= \boldsymbol{I}(U_{ij}(.)\geq c_i)\cdot \boldsymbol{I}(U_{ji}(.)\geq c_j) \qquad \text{for any } i,j \text{ that meet} \\ &\equiv \boldsymbol{I}(f(X_i,X_j,u_{ij};\beta)>0) \\ &\text{where } \boldsymbol{I}(.) \text{ is the indicator function.} \end{aligned}$$

In the second line, we represent the joint choice to be friends with a reduced-form mutual friendship function f. The parameters of this function (β) represent tastes for the observed parameters as well the marginal cost of friendship. u_{ij} is a reduced-form representation of students i and j's unobservable tastes for one another.

[17] We assume that agents do not take existing or future links into account when choosing to form a link. This model of network formation is rudimentary in several dimensions. The decision to form a friendship conditional upon meeting is based solely upon the characteristics of the two students. The network formation literature has more developed theoretical models in which a network is the equilibrium outcome of a noncooperative game. For a good survey of the literature on the theory of network formation, see Jackson (2006).

Motivated by our findings in Section 2.3, the characteristics in X that affect the mutual friendship function are race, parental education, SAT score, and political orientation. The functional form used in the simulation is given by:

$$\begin{aligned}
f(X_i, X_j, u_{ij}; \beta) = \\
&\beta_0 + \beta_{\mathrm{WW}}\boldsymbol{I}(\mathrm{race}_i = \mathrm{race}_j = \mathrm{white}) + \beta_{\mathrm{BB}}\boldsymbol{I}(\mathrm{race}_i = \mathrm{race}_j = \mathrm{black}) \\
&\quad + \beta_{\mathrm{HH}}\boldsymbol{I}(\mathrm{race}_i = \mathrm{race}_j = \mathrm{hispanic}) + \beta_{\mathrm{AA}}\boldsymbol{I}(\mathrm{race}_i = \mathrm{race}_j = \mathrm{asian}) \\
&\quad + \beta_{\mathrm{par_edu}}\boldsymbol{I}(\mathrm{parent_edu}_i = \mathrm{parent_edu}_j = \mathrm{both_coll}) \\
&\quad + \beta_{\mathrm{cons}}\boldsymbol{I}(\mathrm{conservative}_i = \mathrm{conservative}_j = 1) \\
&\quad + \beta_{\mathrm{skill}}\boldsymbol{I}(\mathrm{SAT}_i > 1200\ \&\ \mathrm{SAT}_j > 1200) + u_{ij}
\end{aligned}$$

where u_{ij} captures the joint effect of the unobservable characteristics of i and j. The β coefficients capture tastes for similar characteristics. In the calibration of the model, u_{ij} is simulated with independent random draws from a normal distribution.[18] The mean and variance are normalized to zero and one. The magnitudes of the other parameters in the function are relative to the variation in the random component.

The meeting is modeled in different stages. First each student meets every other student in the university with probability p_{init}, and this probability is chosen to generate on average c_{init} meetings per person. Next each student meets each other student from the same college with a probability $p_{i\mathrm{COLL}}$, chosen to generate an average of c_{COLL} meetings per person.[19] Students of the same cohort meet each other with probability p_{Year}. Students living in a dorm meet each other student living in the same dorm with probability p_{DORM}. The rule $\boldsymbol{I}(f(X_i,X_j,u_{ij};\beta)>0)$ is used to decide whether a meeting through any one of these channels results in a friendship.

After meeting a set of initial friends, students meet the friends of their friends. This process is motivated by the clusteredness of the networks documented in Section 2. A model with different probabilities of friendship formation can generate the segmentation observed, but cannot produce clusteredness within subgroups.[20] The process of meeting friends of friends can magnify any effects of the institutional environment on friendship formation. Each student meets each friend of her friends with probability p_{frofr}[21]. The friend of friend meeting process is repeated S times. Again $\boldsymbol{I}(f(X_i,X_j,u_{ij};\beta)>0)$ is used to decide whether meeting results in a friendship.

These multiple rounds of meeting and consequent decision whether to form a friendship result in a friendship matrix ***g***. We calculate features describing this simulated network. In Section 4, we calibrate the 14 parameters to fit 14 moments of the simulated network to 14 moments of the actual network at Texas A&M. The moments are: the mean, variance and skewness of the number of friends; the cluster coefficient; the fraction of friends from the same college, same dorm and same cohort; the fraction of friends who are the same race for Whites, Hispanics, Asians and Blacks; the fraction of friends who are high SAT scorers for high SAT scorers; the fraction of friends of the same parental education level; and the fraction of conservative friends of conservatives.

The mechanics of the model implies that all parameters of the model affect all moments. But it is possible to illustrate how the moments are determined by describing the relationship between the different parameters and a given moment. The number of total friends is directly related to the number of students randomly met. The channel of meeting friends of friends generates the variance and skewness of the distribution of the number of friends, as well as,

[18] To check whether the results are sensitive to the independence assumption, we recalibrate the model while imposing a connection between u_{ij} and u_{ik}, and u_{jk} Details can be found in the working paper. Allowing for correlated preferences leads to a larger role for the friends of friends channel. For the counterfactuals we perform below, relaxing the non-correlation assumption has little effect on the resulting network. The only exception is the "special introduction" counterfactual (described below) where assuming independence overstates the effect of the policy on segmentation. Therefore, allowing for correlated preferences does not change our conclusion that university policies have limited ability to reduce observed segregation.

[19] Texas A&M has 10 different academic colleges, e.g. Liberal Arts, Engineering, or Architecture. Because some colleges are larger than others and a student is less likely to meet any other student in the college if the college is large, we allow the probability p^i_{COLL} to vary by individual. The probability varies in such a way that every student meets c_{COLL} students on average from their college, independent of the size of their college.

[20] If both i and j are friends with k, the friends of friend mechanism increases the probability that i meets j. A closed triangle between i, j, and k contributes to clusteredness. The meeting of friends of friends also generates the positive degree correlation.

[21] We relax this assumption by allowing the number of friends who are met in each cycle to depend on the number of the individual's current friends (see the working paper). The counterfactual simulations based on the alternative specification generate very similar results as in the original specification.

the clusteredness. Hence these three moments are directly related to the number of cycles of meeting friends of friends (S), the probability of meeting friends in each cycle (p_{frofr}), and the probability of forming a friendship conditional on meeting, captured by the intercept (β_0) in the $f(.)$ function. The fraction of friends in a similar environment is directly related to the probabilities of meeting people in that environment (c_{COLL}, p_{YEAR}, p_{DORM}). The fraction of friends with the same characteristics implies values for the importance of sharing these characteristics when deciding to form a friendship, i.e. the parameters β_{WW}—both White, β_{BB}—both Black, β_{AA}—both Asian, β_{HH}—both Hispanic, β_{HiSAT}—both high SAT score, $\beta_{\text{par\_edu}}$—same parental education, and β_{cons}—both conservative.

3.2. Assumptions and exclusion restrictions

The model postulates that the probability that two students meet is determined by specific institutional factors (academic college, dorm, cohort). Preferences for friendship conditional upon meeting are determined by specific observable characteristics (race, parental education, political orientation, and academic ability). We assume that unobserved determinants of tastes are uncorrelated with institutional meeting channels. If this assumption is violated, the model will yield biased parameters of the effects of institutional variables. For example, suppose two political science majors share an (unobserved) interest in campus politics. The shared major is an institutional meeting channel that is correlated with the unobserved shared taste for politics, and therefore affects the probability of becoming friends conditional upon meeting. If two political science majors are more likely to be friends conditional upon meeting, we bias upwards the parameters that capture the causal effect of sharing the same major.

Similarly, we assume that unobserved determinants of meeting are uncorrelated with observable taste characteristics. For example, we rule out that two high achieving students in the same dorm are more likely to meet through unobserved institutional meeting channels than a high and a low achieving student in the same dorm. In particular, this assumes that the university does not have unobserved meeting channels that affect the probability of meeting but are correlated with our measures of taste (e.g. honors classes, student associations for certain ethnicities).

These assumptions are motivated by the reduced-form regressions in Section 2. The coefficient estimates of the institutional variables (e.g. same college, cohort) are fairly robust to the inclusion of a variety of covariates on ethnicity, family background, and ability. If the additional covariates pick up any of the unobserved heterogeneity, the robustness of these regressions suggests that the bias may not be severe. In addition, the coefficient estimates of same race are robust to the inclusion of institutional variables. This suggests that the observed institutional variables are related to friendship in a manner that is largely independent of race, and supports the validity of our key identifying assumption.

4. Results and simulations

4.1. Model calibration

We calibrate the 14 parameters of the model to fit 14 moments of the simulated network to 14 moments of the actual network at Texas A&M.[22] We calculate the data moments using the Facebook network of 1930 students randomly drawn from the sample introduced in Section 2.3.[23] The 14 data moments are displayed in column one of Table 4. We use the characteristics of the 1930 individuals and simulate a network by applying the network formation mechanism presented in Section 3.

We choose the parameters to minimize the difference between the features of the simulated network and the features of the network at A&M. The simulations are based on random draws for who meets whom and the elements of u_{ij}. For each set of parameters, the network features of the simulated model are calculated by averaging over 100 simulated

[22] We do not estimate but calibrate the model, which means that the resulting parameters cannot be used for testing or to construct confidence intervals.

[23] Due to computational limits, we have to restrict ourselves to a subset of the 7719 students in the full sample. Therefore, the reader should interpret the counts we present below as corresponding to a “scaled down” network. We drew different samples of 1930 students. The characteristics of the network remain essentially unchanged. While students with many friends will loose more friends due to sampling than students with few friends, the relative distribution of the number of friends is not affected by using a random sub-sample.

Table 4
Parameters of the model under the calibration and outcomes of the counterfactual experiments

Counterfactuals

	(1) Sample of 1930 students	(2) Full model simulation	(3) Completely random friends	(4) Full model without friends of friends	(5) Random meeting	(6) No preferences	(7) Affirmative action, double Hispanics	(8) Introduction to students of different race
Parameters								
S cycles of meeting friends of friends	–	8	0	0	8	8	8	8
c_{init} average no. randomly met	–	6.15	6.41	14.58	25.56	6.16	5.70	4.98
p_{frofr} probability of meeting friends of friends	–	0.54	0	0	0.54	0.54	0.54	0.54
c_{COLL} average no. met same college	–	4.60	0	10.90	0	4.60	4.26	3.73
p_{Year} probability of meeting in same year	–	0.02	0	0.05	0	0.02	0.02	0.02
p_{Dorm} probability of meeting in same dorm	–	0.35	0	0.83	0	0.35	0.32	0.28
β_{Const} constant	–	−1.72	0	−1.72	−1.72	−1.57	−1.72	−1.72
β_{WW} (Whites)	–	0.07	0	0.07	0.07	0	0.07	0.07
β_{BB} (Blacks)	–	2.10	0	2.10	2.10	0	2.10	2.10
β_{HH} (Hispanics)	–	0.40	0	0.40	0.40	0	0.40	0.40
β_{AA}(Asians)	–	0.85	0	0.85	0.85	0	0.85	0.85
β_{skill}(high SAT)	–	0.10	0	0.10	0.10	0	0.10	0.10
$\beta_{\text{par\_edu}}$ (parental education)	–	0.09	0	0.09	0.09	0	0.09	0.09
β_{cons} conservative	–	0.12	0	0.12	0.12	0	0.12	0.12
Network moments								
Average no. of friends	6.42	6.42	6.41	6.42	6.41	6.42	6.41	6.41
Variance of no. of friends	6.44	6.27	2.52	2.96	5.56	5.77	6.40	6.14
Skewness of no. of friends	1.82	1.82	0.39	0.69	1.58	1.56	1.88	1.79
Cluster coefficient	0.15	0.16	0.00	0.01	0.17	0.16	0.16	0.17
Fraction from same year	0.44	0.44	0.25	0.59	0.25	0.44	0.45	0.39
Fraction from same college	0.22	0.22	0.13	0.31	0.13	0.21	0.21	0.20
Fraction from same dorm	0.08	0.07	0.01	0.14	0.01	0.07	0.08	0.06
Fraction White friends of Whites	0.87	0.85	0.82	0.85	0.85	0.82	0.76	0.77
Fraction Hispanic friends of Hispanics	0.21	0.22	0.12	0.23	0.22	0.12	0.42	0.21
Fraction Asian friends of Asians	0.15	0.14	0.04	0.14	0.14	0.03	0.12	0.14
Fraction Black friends of Blacks	0.32	0.33	0.02	0.22	0.28	0.02	0.28	0.31
Fraction high SAT score friends of high SAT	0.49	0.49	0.39	0.47	0.47	0.41	0.47	0.48
Fraction friends of same parental education	0.53	0.53	0.44	0.50	0.52	0.45	0.50	0.51
Fraction conservative friends of conservative	0.62	0.62	0.52	0.59	0.61	0.53	0.60	0.60

Notes:
(1) The data used are a random sample of 1930 of the 7719 students described in Section 2.3.
(2) Are the parameters of the full model calibration that fit the simulated moments to the moments of the actual network.
(3) Students meet with the same probability independent of school environment, they do not have preferences for characteristics and do not meet friends of friends.
(4) Students meet with probabilities that vary with the school environment, they have preferences for characteristics, but they do not meet friends of friends.
(5) Students meet with the same probability independent of school environment, they have preferences for characteristics, and they meet friends of friends.
(6) Students meet with probabilities that vary with school environment, they do not have preferences for characteristics, and they meet friends of friends.
(7) Double the number of Hispanic students (with parameters of full model but meeting probabilities scaled down to generate the same average number of friends).
(8) Add an extra meeting round where each White meets 1% of minority students and each minority student meet 1% of White students (with parameters of full model but meeting probabilities scaled down to generate same average number of friends).

networks with different random draws. The 100 sets of random components are kept constant for each different set of parameters. The resulting parameters are displayed in column 2 of Table 4.

The simulated network is generated by meeting on average 6.1 random students, 2.1% of all students in the same cohort, and 4.6 students from the same college. Students living in the same dorm meet each other with 35% probability. Conditional on meeting, Whites have a very small preference for friendships with other Whites (β_{WW}=.07). The preferences for same race friendships are much stronger for Hispanics (β_{HH}=.40) , Asians (β_{AA}=.85), and especially Blacks (β_{BB}=2.10). The preferences for friends with similar SAT scores, parental background or political orientation are less pronounced than the preferences for same race friendships among minorities.

The lower part of Table 4 displays the features of the network simulated with these parameters. Our model generates the features of the network. It matches the average number of friends, the variation in the number of friends, the right skewed distribution of the number of friends, and the clusteredness.[24] Also, it matches the likelihood of forming a friendship for students sharing a similar environment or similar characteristics.[25]

4.2. Counterfactual experiments

To simulate social networks under counterfactual policies, we use the parameters obtained above but change various elements of the network formation process. We assume that the parameters of the model are not affected by the policy changes. This can be justified by the fact that any actual changes are most likely only marginal. However, for more substantial policy changes our approach is subject to the Lucas critique.

A benchmark is purely random friendship formation. Each student meets each other independent of their environment, and the probability of forming a friendship does not depend on any characteristics of the students. Under purely random friendships, the fraction of friends with certain characteristics reflects the share of the total population, and there is no segmentation. The features of the resulting network are shown in column 3 of Table 4.

The first counterfactual experiment shows that meeting friends through other friends can magnify certain measures of segmentation and mitigate others. In this simulation, we “turn off” the friends of friends meeting channel; students meet with probabilities that vary in school environment and they have preferences for friend characteristics, but they do not meet friends of their friends. The parameters for this counterfactual simulation and the resulting network features are shown in column 4 of Table 4.[26] The segmentation based upon dorm and cohort is larger, suggesting that the friends of friends channel facilitates becoming friends with students in other school environments. However, the friends of friends channel magnifies racial segmentation for Blacks −22% of the friends of Blacks are Black without the friends of friends channel while 33% are Black when we allow for meeting friends through friends. This result illustrates that accurately modeling the effects of environmental changes needs to incorporate the feature that two students being friends will be a function of their other friends.

Column 5 shows the parameters and resulting network features for the counterfactual of “random meeting”. This counterfactual models the extreme case of a university eliminating any meeting channels that generate segmentation. Obviously, it would be impossible to eliminate all such channels, but a university could, for example, create a common set of core classes that students from all academic colleges must take. In this simulation, each student meets every other student with equal probability (i.e. the institutional meeting channels do not affect meeting probabilities), but students have the preferences that we estimate above and meet the friends of their friends.[27] We find that the variation and skewness of the number of friends decrease slightly but the cluster coefficient remains virtually unchanged. As expected, the disproportionate number of friends with a similar campus environment disappears. However, the segmentation by race, ability, political orientation and parental education largely persists. This is a potentially sobering

[24] The features are not perfectly fitted. One reason is that the number of cycles of meeting friends of friends is an integer rather than a continuous parameter.

[25] To assess the suitability of the model, we compare other moments of the actual network (not used in the calibration) to comparable moments of the simulated network. While not all moments fit in exact magnitude, the model is able to replicate the general data patterns. See the working paper for details.

[26] We scale up the meeting probabilities to generate the observed average number of friends.

[27] Mechanically, the average number of random encounters of each student is picked to generate the same average number of friends as in the original network.

Table 5
Associations between student outcomes and peer characteristics

Dependent variable:	Own GPA		Drinker	Volunteer		Religious		Political	
				Excl. Same Orgs	Incl. Same Orgs	Excl. Same Orgs	Incl. Same Orgs	Excl. Same Orgs	Incl. Same Orgs
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Own characteristics									
SAT in 100s of points	0.096	0.093	−0.006	0.002	0.002	0.004	0.003	−0.005	−0.006
	[0.007] **	[0.007] **	[0.004]	[0.004]	[0.004]	[0.004]	[0.003]	[0.023]	[0.023]
High school percentile (0–100)	0.015	0.015	−0.001	0.001	0.001	0	−0.001	0.002	0.002
	[0.001] **	[0.001] **	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.003]	[0.003]
Female	0.135	0.117	−0.032	0.061	0.055	0.016	0.015	−0.464	−0.465
	[0.017] **	[0.016] **	[0.010] **	[0.010] **	[0.010] **	[0.009]	[0.009]	[0.058] **	[0.058] **
Parent HH income $40–80 k	−0.039	−0.037	0.016	0.007	0.007	−0.009	−0.009	0.003	0.004
	[0.025]	[0.025]	[0.014]	[0.015]	[0.015]	[0.013]	[0.013]	[0.086]	[0.086]
Parent HH income >$80 k	0.009	0.008	0.031	0.006	0.002	−0.016	−0.014	0.004	0.003
	[0.026]	[0.026]	[0.014] *	[0.015]	[0.015]	[0.014]	[0.013]	[0.088]	[0.088]
Father college graduate	0.068	0.060	−0.008	0.003	0.001	0.024	0.019	0.097	0.098
	[0.018] **	[0.018] **	[0.010]	[0.011]	[0.010]	[0.010] *	[0.009] *	[0.061]	[0.061]
Mother college graduate	0.015	0.015	0.002	0.009	0.01	0.006	0.006	−0.05	−0.049
	[0.016]	[0.016]	[0.009]	[0.010]	[0.010]	[0.009]	[0.009]	[0.057]	[0.057]
High school percentile economically disadvantaged	−0.003	−0.003	0.000	0.000	0.000	0.000	0.000	−0.002	−0.002
	[0.001] **	[0.001] **	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.002]	[0.002]
High school pass rate of standardized TAAS test	0.006	0.005	0.000	0.001	0.000	0.001	0.001	−0.009	−0.009
	[0.001] **	[0.001] **	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.005]	[0.005]
Friend characteristics									
Predetermined									
Average of friends' SAT in 100s of points	0.030	−0.032	−0.014	0.006	0.009	−0.004	−0.016	−0.088	−0.089
	[0.016]	[0.017]	[0.009]	[0.010]	[0.010]	[0.009]	[0.009]	[0.060]	[0.059]
Average of friends' high school percentile	0.002	−0.005	−0.002	0.000	−0.001	0.002	0.001	0.005	0.006
	[0.002]	[0.002] **	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.007]	[0.007]
Fraction of friends with college educated father	0.160	0.106	−0.019	0.059	0.028	0.099	0.061	0.034	0.030
	[0.058] **	[0.057]	[0.032]	[0.040]	[0.039]	[0.036] **	[0.035]	[0.231]	[0.231]
Fraction of friends with college educated mother	0.029	−0.007	0.032	0.049	0.038	0.036	0.046	0.498	0.494
	[0.057]	[0.056]	[0.032]	[0.039]	[0.039]	[0.036]	[0.034]	[0.228] *	[0.228] *
Contemporaneous									
Average of friends' GPA		0.461							
		[0.044] **							
Fraction of friends who 'drink'		−0.123	0.386						
		[0.075]	[0.042] **						
Fraction of friends in volunteer groups				−0.036	0.564				
				[0.045]	[0.042] **				
Fraction of friends in religious groups						0.497	0.781		
						[0.033] **	[0.028] **		
Fraction of friends in political groups								0.200	0.206
								[0.418]	[0.412]
Constant	−0.612	−0.456	0.573	−0.199	−0.222	−0.306	−0.081	3.789	3.76
	[0.230] **	[0.230] *	[0.140] **	[0.153]	[0.150]	[0.140] *	[0.133]	[0.887] **	[0.886] **
Observations	5138	5138	5138	4655	4661	4655	4661	4655	4661
R^2	0.31	0.33	0.04	0.03	0.07	0.07	0.16	0.04	0.04

Notes: Models estimated via least squares with Huber–White robust standard errors reported in brackets. The dependent variable in the last 4 sets of models is a dummy variable for whether the student is a drinker or a member of a volunteer, religious, or politic organization. All regressions include dummies for race, year in college, academic college on campus (e.g. Liberal Arts or Engineering), the number of members in the student's family household, and the Spectral Segregation Index (SSI) for each race. The sample includes all students among 7719 (used in the model in Sections 2 and 3) who have friends among the 7719 to compute average friend characteristics. "Excl. Same Orgs" and "Incl. Same Orgs" means that the calculation of the fraction of friends in the type of organization either *excludes* or *includes* friends who are in the same organization.

* Significant at 5%.

** Significant at 1%.

result for university administrators. It suggests that university policies geared towards increasing the encounters between different groups of students have very limited ability to reduce segmentation in their students' social networks.

Next, we perform the "reverse" counterfactual—the case of undiscriminatory preferences with existing meeting channels. The likelihood of forming a friendship conditional on meeting does not depend on the characteristics of a person. The probabilities of meeting other students are the same ones as in the full model (i.e. the institutional channels affect meeting probabilities and students meet friends of friends). The parameters we use and results are in column 6. The numbers of same race friendships are very close to those that would arise from purely random friendship formation. This confirms the result that the segmentation according to race in the actual network is mainly driven by preferences and not by different meeting probabilities. This supports that the reason that, say, Hispanics have disproportionately many Hispanic friends is preferences. It argues against the alternative explanation that Hispanics meet disproportionately many Hispanics through channels such as major or dorm and these differences are then magnified through introduction to friends of friends.

We also can simulate the effect of an affirmative action policy that admits more students of a certain demographic profile. We model the admission of more Hispanics who are a large and growing population in Texas. In column (7) we simulate the policy experiment of doubling the population of Hispanic students. We do so by including each Hispanic student with all his/her characteristics twice in the simulation. We assume that preferences for race do not change. We find that Hispanics would have a much more racially segmented social network—the share of Hispanic friends of Hispanics nearly doubles. However, the share of friends of a different race increases for Whites, Asians and Blacks. In particular, the share of non-White friends of Whites increases from 15% to 24%. This implies that increasing the number of Hispanics would lead to modest increases in the racial diversity of interaction for non-Hispanics.

Our final policy counterfactual is to introduce Whites and minority students to each other. This corresponds to intentional efforts by the university to facilitate interaction between students of different backgrounds (e.g. targeted introductions during orientation week). To perform this simulation, we include an extra meeting round, where each White student has a 1% chance of meeting each minority student and each minority student has a 1% chance of meeting each White student. This translates to each White student meeting 3.5 non-White students and each minority student meeting 15 White students. The probability of forming a friendship conditional on meeting is still given by the preferences used to simulate the full model. Given our previous finding that preferences significantly affect friendship formation, we would expect this policy to have only limited effects. Column 8 shows the results. Indeed, we find that the diversity of social interaction only modestly increases. The share of minority friends of White students increases from 15% to 23%. The total number of friends of minority students increases, but their share of same race friends decreases only slightly.

The counterfactual results suggest that changes in institutional policies have limited potential to increase inter-race interaction. Equalizing the probability of meeting (e.g. random dorm assignment and/or a common set of core classes) only negligibly changes the fraction of inter-race friendships for Whites, Hispanics, and Asians. A policy of targeted introductions of minorities to non-minorities has the largest impact on Whites by increasing the number of non-minority friends, but has little impact on Hispanics and Asians. However, alternative institutional policies do have modest impacts on Blacks. Targeted introductions increase inter-race friendships from 67% to 69% while equalizing the meeting probability increases the figure to 72%. This result is likely driven by the finding that the magnification effect of preferences through friends of friends is strongest for Blacks. Thus, policies that change meeting probabilities can modestly impact Blacks' social network despite strong same-race preferences.

5. Evidence that Facebook social networks are associated with educational outcomes

As pointed out in the Introduction, social contacts may affect many outcomes that are of interest to economists. When using data from an educational setting, many of these outcomes are not observable or have not manifested themselves yet. We do not observe how diverse interaction promotes the diversity of thought, how students' social networks affect their job search, or how their eventual productivity is affected by their connections. However, we are able to document some associations of academic and social outcomes (Foster (2006) points out that in student populations, peer effects tend to be more robust in social outcomes than in academic outcomes.[28]) Because we measure both endogenously and exogenously determined friends, we do not interpret any of these associations as causal.[29]

[28] Stinebrickner and Stinebrickner (2006) point out that it is not clear where to look for the most influential peers.

[29] See Manski (1993) for a discussion of these identification problems.

We regress a student's outcome on own-pre-treatment characteristics, friend pre-treatment characteristics, and contemporaneous friend characteristics. We study five student outcomes — GPA, drinking behavior, and participation in three types of organizations. GPA is a student's contemporaneous college GPA. 'Drinker' as an indicator that the student self-reported on their Facebook profile's "Interests" section words typically associated with drinking.[30] We use Facebook profile information on "Jobs and Clubs" to identify students who participate in three types of campus or community organizations — volunteer, religious, and political.[31]

A student's own pre-treatment characteristics include measures of ability (SAT score and high school percentile), family characteristics (parental income and education), high school characteristics, race, and gender. A student's friends' pre-treatment characteristics are average friend SAT score, high school percentile, and parental education. We also include a measure of each student's own-race segregation – the Spectral Segregation Index (SSI) – recently developed by Echenique and Fryer (2007).[32] Finally, we include several measures of the average contemporaneous characteristics of a student's friends.

Results are reported in Table 5. The last group of variables in each regression is the contemporaneous friend characteristics. If these characteristics are strongly related to a student's outcome, as we find below, this suggests that a student's outcomes are strongly related to the outcomes of her friends.

The first 2 columns report the relationships between a student's GPA and her friends' average characteristics. In column 1, we include the student's own characteristics and friends' predetermined characteristics. Students with higher GPAs are those with higher measures of ability, a college educated father, those who went to wealthier high schools with higher standardized tests scores, and those with friends who have college educated fathers. Friends' ability measures – SAT and high school percentile – are not individually significant, but both are positively related to GPA and jointly significant at the 5% level. Column 2 adds measures of contemporaneous friend characteristics. Average friend GPA is very strongly associated with own GPA – an increase in one letter grade in average friend GPA is associated with almost half a letter grade increase in own GPA. Friends' predetermined skill measures are no longer positive after controlling for friend GPA.

Column 3 reports estimates of a linear probability model of being a 'drinker'. Males from higher income families are more likely to be 'drinkers'. Students with friends with lower SAT scores and lower high school percentiles are more likely to be drinkers; the coefficients are not individually significant but are jointly significant at the 1% level. Finally, a student with more friends who are 'drinkers' is substantially more likely to be a 'drinker'. A one standard deviation increase in the fraction of drinker friends (10%) increases the probability of being a 'drinker' by 4%.

The final sets of columns investigate participation in volunteer, religious and political organizations with a linear probability model. For each outcome, we regress an indicator variable for whether the student is a member of the type of organization on own/friend characteristics and the fraction of her friends who are in that type of organization. Friends being in the same type of organization reflect at least two mechanisms: friends may have similar preferences to participate in certain activities, or friends may meet via the organization. In the first column of each model, we calculate the fraction of friends by excluding friends who are in the same organization. This is intended to avoid counting friends who meet through the organization, and primarily identify friends with similar preferences to participate in certain types of organizations. In the second column of each model, we include friends who are in the same organization, so this measure also picks up friends who may meet in the organization.

Column 4 shows that students with more friends in other volunteer groups are no more likely to volunteer, but column 5 shows that having more friends in any volunteer group makes the student much more likely to volunteer. This suggests that volunteer groups serve as a meeting channel.

The association between own and friend membership is even stronger for religious organizations. Even when we include only friends who are in different religious groups, a student with more friends in religious organizations is associated with a substantial increase in the likelihood of being in such a group oneself. This suggests that students who

[30] For example, students who list an Interest in "beer", "liquor", "drinking", and "partying" are classified as a 'drinker'. 10% of students in our sample are 'drinkers' according to this measure.

[31] To measure if a student participates in a these types of organizations, we collected all self-reported organizational membership in the Facebook data. Then, two research assistants who are familiar with student organizations independently classified the organizations as being volunteer, political, religious, or other. We classify the organization if both research assistants agreed on the classification. According to this measure, the percentage of students participating in volunteer, religious and political organizations is 9%, 8%, and 3%, respectively.

[32] The SSI allows researchers to calculate an individual's own-race segregation. Computer code for calculating the SSI can be found at:http://www.economics.harvard.edu/faculty/fryer/projects.html. We thank Kimon Ioannides for his assistance with calculating the SSI.

are friends are likely to have similar underlying preferences for religious fellowship. When we include friends in any religious organization in column 7, the association is even stronger.

For political organizations, the association between own and friend membership is weak. Under both measures of friend membership, an increase in friend membership is associated with higher own membership but the relationship is not statistically significant.

We emphasize that these relationships should not be interpreted causally. A variety of factors – selection into friendships, the reflection problem and endogenous and exogenous peer effects – make causal inference problematic. Nevertheless, we obtain estimates of associations between student outcomes and friend characteristics that are consistent with the existence of some peer effects. Also, these results provide further evidence that Facebook friends are students' peers with whom they interact in class, student organizations, the dorm or other campus activities.

6. Conclusions

Data from the Facebook networks offer insights into the social networks that impact learning, information transmission, and labor market outcomes at the beginning of adulthood. The data provide a large-scale view of the social networks at universities of various sizes. These social networks exhibit many of the characteristics suggested by the network structure literature — clustering, positive degree correlation, and variance and skewness of the degree distribution. In addition, we quantify segmentation along racial and socioeconomic lines and document the diversity of interaction on university campuses.

Our model provides a methodology to analyze segmentation in social networks and decompose the contribution of both school environment and preferences to observed segmentation. Our findings offer a mixed message for university administrators who seek to create diverse social interaction on campus. On one hand, social networks exhibit only modest segmentation across some important dimensions. In the actual network, the fraction of friends with similar ability, parental education, and political orientation does not differ substantially from the fraction that would be generated by random assignment of friends. This suggests that diverse interaction does occur.

However, social networks are highly segmented by race, and this is present at schools ranging from small private institutions (Rice, SMU) to large public universities (Texas A&M and University of Texas). Moreover, our counterfactual simulations suggest that racial segmentation is largely driven by preferences rather than institutional features that affect meeting. Changes in university policies that affect student meeting channels appear to have limited ability to reduce racial segmentation. Therefore, policies aimed at increasing social integration need to somehow impact preferences.

Our network formation model does not allow for strategic interaction between agents. Rather, it provides a purely mechanical meeting process, which is assumed to be independent of preferences for friendship formation. The challenge for future work is to develop models that allow for more complex mechanisms of network formation, while still being empirically tractable.

References

Babcock, Philip, 2006. From Ties to Gains? Evidence on Connectedness, Skill Acquisition, and Diversity, mimeo, UC Santa Barbara.

Bok, Derek C., Bowen, William G., 1998. The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions. Princeton University Press.

Calvo-Armengol, Antoni, Jackson, Matthew O., 2004. The effects of social networks on employment and inequality. American Economic Review 94 (3) (June).

Calvo-Armengol, Antoni, Matthew O. Jackson, 2007. "Like Father, Like Son: Network Externalities, Parent–Child Correlation in Behavior, and Social Mobility", mimeo.

Cohen, Lauren, Frazzini, Andrea, Malloy, Christopher, 2007. The Small World of Investing: Board Connections and Mutual Fund Returns. NBER Working Paper.

Echenique, Federico, Fryer, Roland, 2007. A measure of segregation based on social interactions. Quarterly Journal of Economics 122 (2) (May).

Gigi, Foster, 2006. It's not your peers, and it's not your friends: some progress toward understanding the educational peer effect mechanism. Journal of Public Economics 90.

Fryer, Roland, Torelli, Paul, 2006. An Empirical Analysis of 'Acting White', mimeo, Harvard.

Grutter v. Bollinger, *United States Supreme Court* (02-241) 539 U.S. 306 (2003).

Goyal, Sanjeev, van der Leij, Marco J., Moraga-Gonzalez, Jose Luis, 2006. Economics: An Emerging Small World? Journal of Political Economy 114 (2).

Granovetter, Mark S., 1973. The Strength of Weak Ties. American Journal of Sociology 78 (6).

Ioannides, Yannis M., Loury, Linda Datcher, 2004. Job information networks, neighborhood effects and inequality. Journal of Economic Literature 42 (4) (December).

Jackson, Matthew O, 2006. In: Blundell, Richard, Newey, Whitney, Persson, Torsten (Eds.), The Economics of Social Networks Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, Chapter 1, Volume 1. Cambridge University Press.

Jackson, Matthew O., Rogers, Brian W., 2007. Meeting strangers and friends of friends: how random are social networks? American Economic Review 97 (3), 890–915 (June).

Joyner, Kara, Kao, Grace, 2000. School racial composition and adolescent racial homophily. Social Science Quarterly 81 (3).

Manski, Charles F., 1993. Identification of endogenous social effects: the reflection problem. Review of Economic Studies 60 (3), 531–542 (July).

Manski, Charles F., 2000. Economic analysis of social interactions. Journal of Economic Perspectives 14 (3) Summer.

Marmaros, David, Sacerdote, Bruce I., 2006. How do friendships form? Quarterly Journal of Economics 121 (1) (February).

Montgomery, James D., 1991. Social networks and labor-market outcomes: toward an economic analysis. American Economic Review 81 (5).

Moody, James, 2001. Race, school integration, and friendship segregation in America. American Journal of Sociology 107 (3).

Newman, Mark E.J., 2003. The structure and function of complex networks. Society for Industrial and Applied Mathematics Review 45.

Pellizarri, Michele, 2004. Do friends and relatives really help in getting a job? London School of Economics, CEP Discussion Paper, vol. 623.

Quillian, Lincoln, Campbell, Mary E., 2003. Beyond Black and White: the present and future of multiracial friendship segregation. American Sociological Review 68.

Sacerdote, Bruce I., 2001. Peer effects with random assignment: results for Dartmouth roommates. Quarterly Journal of Economics 116 (2) (May).

Soetevent, Adriaan R., 2004. Social Interactions and Economic Outcomes, thesis, University of Groningen.

Stinebrickner, Todd R., Stinebrickner, Ralph, 2006. What can be learned about peer effects using college roommates? Evidence from new survey data and students from disadvantaged backgrounds. Journal of Public Economics 90.

Ward, Bryce A., 2004. Distance and Social Capital: Can Isolation be Good?, mimeo, Harvard University.

Weinberg, Bruce A., 2005. Social Interactions and Endogenous Association, mimeo, Ohio State University.

Wasserman, Stanley, Faust, Katherine, 1994. Social Network Analysis: Methods and Applications. Cambridge University Press.