Influencing along Observables: Estimating College Students' Sorting into Peer Groups and Its Educational Impact

Jakina Debnam[1]

This Preliminary Draft: 12/31/2016 (Please contact the author before citing)

Abstract

Using a novel set of behavioral social network data which captures online "friendship" and messaging of undergraduate students at an elite northeastern university, I estimate the role of sociodemographic characteristics – gender, race, first-generation status and citizenship – in peer network formation over time. Using exponential random graph modeling, I find evidence that the role played by sociodemographic characteristics – gender, race, first-generation status and citizenship – in network formation when students' relationships with one another are exclusively virtual differs from the role played by these characteristics when relationships may include a face-to-face component (e.g. when students are on campus).

JEL Codes: I2, I24, L15

I. Introduction

The importance of peer effects for educational and labor market outcomes is well documented. For post-secondary education, the literature examining the importance of peer effects for academic outcomes finds robust effects for academic outcomes such as grades and large effects for social behaviors such as smoking[2]. Given the importance of such peer effects, there exists relatively little work on how such post-secondary educational peer networks form in the first place.

[1]Charles H. Dyson School of Applied Economics and Management, Cornell University, 410 Warren Hall, Ithaca, NY 14853 (e-mail: jrd294@cornell.edu). I am grateful to the Behavioral Economics Research Group and Social Dynamics Laboratory at Cornell University, as well as to the National Science Foundation Graduate Research Fellowship Program grant no. DGE-1144153.

[2] For a recent review of the literature concerning peer effects in education, see Sacerdote (2011).

Learning how such peer networks form, and how institutional features interact with this network formation, will better equip administrators to create institutional environments which encourage peer networks that are more conducive to better educational and labor market outcomes, particularly for groups for whom the lack of relevant peer networks can represent a barrier to success.

To provide insights about peer network formation in post-secondary educational environments, I use a measure of social connectedness within a private social network used by cohorts of students attending a particular college ("the College") within a large elite private northeastern university ("the University"), matriculating between the years 2015 and 2020. Analysis is based on behavioral social network and administrative data collected from all undergraduate students. I use a binary relationship of 'followership' within the platform, a measure akin to followership on Twittter but with a key distinction – unlike in the case of Facebook- or Twitter-based relationship measures followership in Chatter serves primarily a signaling function – students do not frequently communicate with one another using the platform, or receive information from followership relationships with other students. Instead students choose to 'follow' people with whom they have connected in real life. I substantiate this assumption using a survey of undergraduate students in the College about their use of the platform.

Using these 'followership' relationship signals, I find that connections are significantly more likely between students who are of the same gender and ethnicity while these students are on campus, however in the period between being admitted to the College and arriving on campus, observable characteristics play a smaller role in peer network formation.

II. Data and Empirical Strategy

Analysis comes from de-identified administrative data collected from all undergraduate students attending a particular the College within the University, between the years 2015 and 2020. Population summary statistics are described in Table 1. The demographic composition of the subset of students for whom 'followership' relationships are observed on the private social network is roughly comparable to the entire population of users of the Chatter platform (i.e. all students in the College).

Table 1: Population Summary Statistics of Students Matriculating 2015 - 2020

	All Users *N=1,928*	*Students who are part of at least 1 following relationship* *N=312*
Gender		
Female	1,042 (56.45%)	174 (55.77%)
Male	803 (43.50%)	138 (44.23%)
Ethnicity		
Asian	193 (10.09%)	48 (14.72%)
Black	171 (8.94%)	30 (9.20%)
Hispanic	93 (4.86%)	21 (6.44%)
White	910 (47.59%)	150 (46.01%)
Other	545 (28.50%)	77 (23.62%)
Citizenship		
U.S.	1,841 (95.49%)	308 (94.48%)
Other	87 (4.51%)	18 (5.52%)
Matriculation Year		
2015	465 (24.12%)	27 (8.28%)
2016	478 (24.79%)	54 (16.56%)
2017	511 (26.50%)	139 (42.64%)
2018	474 (24.59%)	106 (32.52%)
First Gen. Student	120 (6.22%)	49 (15.03%)
Mean GPA	3.26	3.35

Note: Totals may not sum due to missing/inconsistent demographic information for some students.

Empirical analysis of peer network formation exploits novel data from a private social network used by all students at the College – *Chatter*. Chatter is a Salesforce-based platform used by the College to disseminate information to students concerning topics such as registration, class selection, extracurricular activities, and job search, with the self-described objective "to create a forum to assist with course enrollment, orient students to their major, decompress the Fall experience, and build a sense of community among students, faculty, and staff at the College". As shown in Figure 1, usage by students peaks during freshman year when they use the platform to seek information about orientation and registration, and during later years during students' job and internship searches. When implemented, Chatter was explicitly designed to streamline into a single source administrative communications which may have previously taken place via email or snail mail. Therefore, particularly during the time period before students at the College arrive on campus (during which time their relevant external peer networks are plausibly non-existent and therefore a negligible source of information), this platform is a substantial contributor to students' total set of information regarding administrative and academic choices at the university.

However, there are a relatively small number of private direct messages being transmitted across the network. Instead, most users passively receive content through feeds, based on the groups to which they belong and the postings of the individuals whom they choose to *follow*. Electing into a *followership* relationship is a one-way relationship akin to a followership relation on the Twitter platform – it enables students to passively review the content generated by the individual they have chosen to follow. All users are by default members of a "Students of the College" group, and a group defined for their major. Students may then opt-in to additional groups based on their professional and extracurricular interests. Some groups are open to all students, others are invitation-only. As a result, the vast majority of communication within the network takes a one to many form, where administrators broadcast pertinent information to large groups of students.

Because of this aspect of Chatter, following administrators provides a key source of information for students. Student-to-student followership relationships are qualitatively different - unlike relationships involving administrators, these relationships do not provide users with access to expected information or resources within Chatter. Instead, these followerships provide an opportunity for peer-to-peer messages and can signal existing friendship relationships outside of Chatter. Therefore I exclude followership relationships where at least one member of the pair is an administrator.

From these followerships I observe a directed graph, represented by an NXN adjacency matrix $\mathbf{G_d} = [g_{ij}]$, where $g_{ij} = 1$ if agent i is followed by agent j. For analysis, however, I follow conventions of Exponential Random Graph Modeling and collapse this graph into an undirected NXN adjacency matrix $\mathbf{G_u} = [g_{ij}]$, where $g_{ij} = 1$ if either agent i is followed by agent j or agent j is followed by agent i, 0 otherwise. Network-level summary statistics are presented in Table 2. All network instantiations are sparse and display low levels of reciprocity and transitivity.

(Figure 1 Here – Evolution of Chatter Use Over Student-Semesters)

Table 2: Network Summary Statistics of Students Matriculating 2015 - 2020

	All Students in Followership Relationships N = 312
Density	.00463
Transitivity	.0460
Reciprocity	.0298

The network summary statistics detailed in Table 2 illustrate the sparseness of the followership graph (see Appendix for an image of the network graph). For this reason, I include limited network-level controls.

III. Student Survey of Chatter Use

The interpretability of this analysis hinges critically on understanding what a student's choice to 'follow' another student signifies. For this reason, I conducted a convenience survey of 170 undergraduate students at the College to ask them directly about the meaning of their followership choices, and about the informational value (or lack thereof) of these choices (see Appendix for full survey text). While students report using Chatter to get information about social activities (11.02% of respondents), registration (29.06% of respondents), and classes (33% of respondents), this information comes from administrators in the platform rather than from other students. The clear majority of students (n=74 or 62.71% of students who answered this question) reported that the primary way that they receive information from the platform is via their emailed posting digests, rather than through their followership relationships. An additional 10.17% (n=12) of students who answered this question reported that they do not get any information at all from the platform.

IV. Model of Social Connectedness

A cost-benefit model of friendship formation has been suggested to explain why social connectedness is more likely to occur among students of the same race and gender (Marmaros and Sacerdote 2006). One reason why the costs of friendship may differ along observable characteristics is the sociological tendency of homophily. The statistical importance of homophily (or "similarity breeds connection" (McPherson, Smith-Lovin, and Cook 2001)) for network formation is well-documented in sociology and along observable characteristics such as race and gender, as well as along unobservable characteristics such as values and attitudes. This tendency has been observed in closely related work examining the social connectedness of college students. For example, Marmaros and Sacerdote (2006) find that race and residential proximity are important determinants of Dartmouth students' social interaction. Mayer and Puller (2008) find that two college students on Facebook are more likely to form friendships if they are of the same race, major, cohort, and or political orientation. Homophily may arise through individuals' selection of one another based on similarity ('preference'); or through individuals' adaptation to become more like those who are close to them ('influence'); or through the increased likelihood of similar of individuals to interact with one another and to form ties ('propinquity').

Recent work in network analysis has investigated the role of homophily in virtual environments. Huang, Shen, and Contractor (2013) find that for online gaming, offline homophily in age and in geographic space continue to have a robust effect on network formation. Tarbush and Teytelboym (2012) use data from Facebook and find that homophily is operative through the propensity of individuals to be friends with those who occupy similar social position (operationalized by similarity in number of Facebook friends).

V. Empirical Analysis of the Features that Make Individuals More Likely to Follow One Another

To understand the role of students' observable common characteristics in their peer network formation, I include network-level Markov Chain Monte Carlo Maximum Likelihood parameter estimates of Exponential Random Graph Models (ERGMs) and their standard errors for the Chatter followership network for several subpopulations, specifically focusing on first-generation and ethnic minority students. Increasingly used in economics[3] to explain network structure through micro-level individual behavior, these ERGM models specify a "distribution" of networks which display specified network properties, and determine which hypothesized individual-level behaviors make the observed network structure more or less likely. A key advantage of these models is that they do not require the behavioral assumption of independence necessary for logistic regression. Using an undirected dichotomous measure of followership, this analysis explores how similarity along observable dimensions impacts the likelihood of two individuals forming a social connection, and how this likelihood differs before and after students are on campus.

Table 3 reports ERGM estimates the impact of shared gender, race, citizenship, first-generation status and cohort on the likelihood of formation of followership ties in Chatter, estimating a uniform coefficient for each attribute's impact on network formation. The sample has been limited to cohorts enrolling in the University in 2014 and 2015 - those cohorts for which significant amounts of social connections exist in the data both before and after they arrived on campus. Each estimated model includes controls for the density of the network but lacks a control for transitivity, this methodological choice is informed by both the low level of transitivity in all network instantiations and the failure of all models including this term to converge. Column 1 provides parameter estimates for all social connections, Column 2 provides parameter estimates for social connections formed between acceptance to the College and arrival on campus, Column 3 provides parameter estimates for social connections formed after students arrived on campus.

[3] For a discussion see Chandrasekhar and Jackson 2014.

Table 3: Exponential Random Graph Models – Changing Role of Observable Characteristics for Undergrad Cohorts Entering the University in 2014 or 2015

	All 'Followerships'	Pre-Campus 'Followerships'	On-Campus 'Followerships'
	(1)	(2)	(3)
Edges	-6.787***	-6.282***	-5.268***
	(0.314)	(0.400)	(0.475)
Gender	-0.034	-2.597***	0.035
	(0.130)	(0.344)	(0.216)
Ethnicity	-.520**	0.841***	0.178
	(0.163)	(0.161)	(0.235)
Citizenship	0.805**	0.548	0.295
	(0.267)	(0.302)	(0.428)
First-Gen. Status	0.776***	0.678*	0.591*
	(0.183)	(0.281)	(0.243)
Matriculation Year	0.490***	0.601***	0.188
	(0.129)	(0.166)	(0.239)
Model AIC	3,039	1,760	959.5
Residual deviance	3,027	1,748	947.5
(df)	(52,206)	(25,754)	(7,826)

***p<.001. **p<.01. *p<.05.

The coefficients of the above are log-odds, which estimate how much more or less likely a social connection is to form between two individuals who share a given characteristic, controlling for reciprocal tie formation[4]. From column 1, we see that across the cohorts of undergraduates enrolling in the University in 2014 and 2015, having in common any of U.S. citizenship, first-generation status and cohort year make ties more likely to form between individual students. However, we see that having in common the same ethnicity, in general, makes individuals' *less* likely (59.4%) to form ties. When we disaggregate the social connections formed in the pre- and post-campus periods however, we see that this effect is driven by the lack of statistically significant propensity for members of the same ethnicity to form ties in the post-period (shown in Column 3). Social connections in the pre-campus period, which largely reflect social connections made during orientation campus visits, are significantly more likely to occur between students of the same ethnicity (over 231% more likely).

IV.II Differences for Visible and Invisible Students

[4] Reciprocal tie formation ("Edges" in Table 3) is a property of relational networks which has been consistently observed by sociologists. It is best practice to control for this property when estimating ERGMs.

Given that some sorting along observables in taking place in all periods, a natural robustness check follows from the role of these observables' visibility within the platform on the likelihood of tie formation. In an online environment where individuals are less able to perceive the observable characteristics of their peers, selection mechanisms should be mitigated – individuals would be less likely to select peers who are like them when comparability cannot be observed. If followership is merely a reflection of some other peer network, we should not see increasing sorting on observables when these observable characteristics (particularly gender and ethnicity) are saliently featured in a profile photo. In Table 4, I estimate identical ERG models, analyzing separately ties where the individual who was *followed* had uploaded a profile image of themselves, from when they had not.

We observe the opposite empirical fact – the social connections I observe in Chatter are more likely to include sorting on ethnicity and gender if the person being followed did not include a profile image in their Chatter profile. While statistical power is indeed limited by the small number of individuals which have uploaded photos, we see that even point estimates contradict sorting on observables for these populations.

Table 4: Exponential Random Graph Models – Salient Visible Characteristics

	All 'Followerships'	Includes Profile Photo	Excludes Profile Photo
	(1)	(2)	(3)
Edges	-6.787***	-4.267***	-6.729***
	(0.314)	(0.525)	(0.368)
Gender	-0.034	-0.327	0.324*
	(0.130)	(0.170)	(0.156)
Ethnicity	-.520**	0.374	0.557***
	(0.163)	(0.142)	(0.157)
Citizenship	0.805**	-0.144	0.553*
	(0.267)	(0.472)	(0.280)
First-Gen. Status	0.776***	0.388	0.554*
	(0.183)	(0.262)	(0.250)
Matriculation Year	0.490***	0.388*	0.538***
	(0.129)	(0.262)	(0.155)
Model AIC	3,039	742	2,047
Residual deviance	3,027	730	2,035
(df)	(52,206)	(4,154)	(29,750)

***p<.001. **p<.01. *p<.05.

My analysis provides suggestive evidence of a changing role of observable connections in informing the likelihood of tie formation for post-secondary network formation in Chatter over time.

References

Huang, Y., Shen, C., Williams, D., & Contractor, N. (2009, August). Virtually there: Exploring proximity and homophily in a virtual world. In 2009 International Conference on Computational Science and Engineering (pp. 354-359). IEEE.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. "Identification of Peer Effects through Social Networks." Journal of Econometrics 150, no. 1 (May 2009): 41–55.

Chandrasekhar, Arun G., and Matthew O. Jackson. "Tractable and Consistent Random Graph Models." Working Paper. National Bureau of Economic Research, July 2014.

McPherson, M., Smith-Lovin, L., & Cook, J. M.. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*, 415–444. Retrieved from http://www.jstor.org/stable/2678628

Marmaros, David, and Bruce Sacerdote. "How Do Friendships Form?" *The Quarterly Journal of Economics* 121, no. 1 (2006): 79–119.

Mayer, Adalbert, and Steven L. Puller. "The Old Boy (and Girl) Network: Social Network Formation on University Campuses." Journal of Public Economics 92, no. 1–2 (February 2008): 329–47.

Sacerdote, Bruce. "Chapter 4 - Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?" In *Handbook of the Economics of Education*, edited by Stephen Machin and Ludger Woessmann Eric A. Hanushek, 3:249–77.

Sacerdote, B. (2001). Peer Effects with Random Assignment: Results for Dartmouth Roommates. The Quarterly Journal of Economics, 116(2), 681–704.

Tarbush, B., & Teytelboym, A. (2012). Homophily in online social networks. In Internet and Network Economics (pp. 512-518). Springer Berlin Heidelberg.

Zimmerman, David J., 'Peer Effects in Academic Outcomes: Evidence from a Natural Experiment,' Williams Project on the Economics of Higher Education Discussion Paper No. 52, 1999.

APPENDIX

Appendix Survey 1: Survey of Undergraduate Students Concerning Their Chatter Use (*n*=170)

Survey of Social Network Use (7 Questions)

Are you a student in the College? Yes No Don't Know

How many of your close friends do *you follow* on Chatter? _________

How many of your close friends *follow you* on Chatter? _________

If you need information **about social activities**, what social networks do you use to get answers? [PLEASE SELECT ALL THAT APPLY]

- ☐ Academia.edu
- ☐ Chatter
- ☐ Facebook
- ☐ Google Plus
- ☐ LinkedIn
- ☐ Reddit
- ☐ Slack
- ☐ Tumblr
- ☐ Twitter
- ☐ Vine

If you need information **about registration**, what social networks do you use to get answers? [PLEASE SELECT ALL THAT APPLY]

- ☐ Academia.edu
- ☐ Chatter
- ☐ Facebook
- ☐ Google Plus
- ☐ LinkedIn
- ☐ Reddit
- ☐ Slack
- ☐ Tumblr
- ☐ Twitter
- ☐ Vine

If you need information **about classes/coursework**, what social networks do you use to get answers? [PLEASE SELECT ALL THAT APPLY]

- ☐ Academia.edu
- ☐ Chatter
- ☐ Facebook
- ☐ Google Plus
- ☐ LinkedIn
- ☐ Reddit
- ☐ Slack
- ☐ Tumblr
- ☐ Twitter
- ☐ Vine

In what way do you *most frequently* get information from Chatter?

- ☐ View email digests
- ☐ Seek information directly from the website

☐ Receive direct messages from others on the platform

☐ Other (Please specify): ________________

Appendix Graph 1: of Chatter Followership Relations (Including Administrators)