Exploring homophily in demographics and academic performance using spatial-temporal student networks

Quan Nguyen
University of Michigan
quanngu@umich.edu

Oleksandra Poquet
University of South Australia
sspoquet@gmail.com

Christopher Brooks
University of Michigan
brooksch@umich.edu

Warren Li
University of Michigan
liwarren@umich.edu

ABSTRACT

Network analysis in educational research has primarily relied on self-reported relationships or connections inferred from online learning environments, such as discussion forums. However, a large part of students' social connections through day-to-day on-campus encounters has remained underexplored. The paper examines spatial-temporal student networks using campus WiFi log data throughout a semester, and their relations to the student demographics and academic performance. A tie in the spatial-temporal network was inferred when two individuals connected to the same WiFi access point at the same time intervals at the 'beyond chance' frequency. Our findings revealed that students were more likely to co-locate with the individuals of similar gender, ethnic group identity, family income, and grades. Analysis of homophily over the semester showed that students of the same gender were more likely to co-locate as the semester progressed. However, co-location of the students similar on ethnic minority identity, family income, and grades remained consistent throughout the semester. Mixed-effect regression models demonstrated that features derived from spatial-temporal networks, such as degree, the grade of the most frequently co-located peer, and average grade of five most frequently co-located peers were positively associated with academic performance. This study offers a unique exploration of the potential use of WiFi log data in understanding of student relationships integral to the quality of college experience.

Keywords

Network analysis, homophily, spatial-temporal data, WiFi log data.

1. INTRODUCTION

With massification and globalization of higher education, students are exposed to individuals from a different nationality, ethnicity, gender, and socio-economic background. Universities have long been known as physical spaces where students form lifelong social connections, both for professional social capital and personal networks, such as friendship and marriage [1]. Therefore, understanding how social connections form and change in educational settings, as well as the impact student networks have on learning outcomes, can inform educators of unique ways to improve learners' experience [2].

Educational research offers a range of literatures focused on student networks in both face-to-face and blended or online settings [2-9]. Social scientists have conventionally derived student networks from self-report surveys [2, 10]. These surveys ask students to list who they are friends with or who they seek advice from [2]. The data can be collected multiple times to track the changes in network formation [10, 11]. Self-reported networks are a source for much of the extant evidence about student networks. However, such data collection is vulnerable to sampling biases (i.e. a low response rate, a sample from one class) where important network observations may be omitted. The timing of surveys may affect derived network features, and frequent surveying of learners can lead to survey fatigue and a lack of responses.

Instead of self-reports, the EDM and LAK communities have based their network studies on the log-data generated from online discussion forums [3, 4, 12-14]. Digital traces from online discussion enabled researchers to capture the structure, frequency, as well as the content of communication exchanges. Student networks constructed from online logs also have limitations. For instance, many online courses do not require that students use online forums. In face-to-face or blended learning settings, students are also less likely to use discussion forums. Therefore, student networks derived from online communication are limited in their generalizability, which remains a major challenge for researchers in this domain.

One underexplored data source for social network research in educational settings is location-based data. Social scientists have long argued that those in close physical proximity are more likely to form a social connection (McPherson, Smith-Lovin and Cook [1], p.430). More recently, relationship between geographic proximity and social ties have been corroborated by fine-grained geo-location-based analysis using mobile technologies. For example, the Copenhagen Networks Study [15] quantified the impact of physical proximity on student network structures using 500 GPS-enabled smart phones. Eagle, Pentland and Lazer [16] also used mobile technologies to compare the network based on physical proximity with the self-report social network and reported that 95% of the network friendships can be accurately inferred from sensor data. Although student location data from GPS and Bluetooth signals has shown to be informative, such methods are expensive to replicate and challenging to scale due to a high equipment cost.

This paper presents yet another source for location-based data to infer student networks. The paper reports on the study of student networks constructed from routinely collected WiFi logs. Such network data is created transparently to the learners as they connect to campus WiFi access points which are ubiquitous across physical campuses. Spatial-temporal ties between users can be inferred based on the overlap of time intervals in which learners connected to the same access point, suggesting a

Quan Nguyen, Oleksandra Poquet, Christopher Brooks, and Warren Li "Exploring homophily in demographics and academic performance using spatial-temporal student networks" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 194 - 201

reasonably close spatial co-location (room level). The study aims to understand the relationship between spatial-temporal ties and student characteristics across time, and predictive potential of the features derived from spatial temporal networks.

1.1 WiFi network data in education research

Wireless local area networks (WLANs) are ubiquitous in higher education as they provide on-campus Internet access to students, teachers, and staff. Despite extensive research using WiFi data, only a limited number of studies has explored their application for educational purposes [17-20]. A common example is the usage of WiFi data to visualize mobility patterns. For example, the iSpots project showed how people move around campus in real-time [17]. Hang, Pytlarz and Neville [20] combined WiFi logs with information about the buildings to extrapolate user preferences, and to predict user locations using graph embeddings. WiFi data has also been used in predictive modelling. Sarkar, Carpenter, Bader-El-Den and Knight [19] estimated the correlations between students' on-campus time based on WiFi logs and academic performance. Zhou, Ma, Zhang, SuiA, Pei and Moscibroda [18] utilized WLAN data to estimate students' punctuality for lectures to assess the lecture's engagement using mobile phone's interactive states at minute-scale granularity.

An application of WiFi data which has yet to be explored in areas such as EDM is the formation of social network among students on campus. In line with previous research on location-based networks [15, 16], social ties between WiFi users can be inferred from spatial and temporal co-occurrences (i.e. two users connected to the same WiFi access point during the same time window). Compared to surveys, discussion forum data, and proximity data collected through mobile devices (e.g. Bluetooth beacons), WiFi data provides a fine-grained alternative that records the dynamic changes in social interactions over a long period of time. Importantly, WiFi logs can capture physical social interactions and can scale at a relatively low cost. This paper presents initial steps towards exploring spatial and temporal information in the analysis related to student learning.

1.2 Research questions

Individuals are likely to share social connections with others similar to them, a phenomenon known as homophily [1, 21]. In educational settings, researchers have observed homophily based on gender [22], ethnicity [23], international/domestic country of origin [10], study major [24], socio-economic status [23], and academic performance [25, 26]. It might be expected that high-performing students seek friendship with other high-performing peers as part of their academic identity [27, 28], or that groups of high performing learners joined by lower performing learners will raise up those learners [29]. While there has been a large literature exploring the homophily effects in educational settings using traditional questionnaires or interactions in online learning environments, there remains a paucity of research that utilizes location-based data for such purposes. We hypothesize that students with similar traits are more likely to spend more time together on campus, i.e. in a spatial temporal co-occurrence from which a social connection can be inferred.

RQ1: How do demographic characteristics and grades affect the likelihood of spatial-temporal co-occurrence among students?

Second, we examine if spatial-temporal student network can capture social selection processes among students, also a phenomenon previously observed in social student networks. 'Social selection' refers to the choice to interact with others of similar status or value, and has been observed in various educational settings [27, 28]. With the increasing availability of digital data in education (i.e. LMS, online discussion forums), researchers are enabled to observe the dynamics of social selection processes with high temporal precision. In these regards, we are interested in understanding the temporal changes in the homophily effects of demographics and academic performance over time. For example, one might expect that at the beginning of the semester, students are more likely to form friendships based on similarity in demographic attributes as they have not acquired sufficient information about their peers' academic ability. One might also expect that as students approach the end of the semester, more social ties will be formed within similar performance groups. This leads us to our second research question:

RQ2: How does homophily based on demographic characteristics and academic performance change over time?

In addition to these questions, previous studies [7, 21, 26] have confirmed a positive relation between the degree of social integration/participation and academic performance. Motivated by this, we are interested in the predictive potential of 'peer effects' for grade performance using location-based network data. The relationship between that of a peer and one's characteristics has been studied for dormmates, as well as classmates, schoolmates, or children from the same neighborhood [29]. Administrative records of class co-enrolment have also been shown to capture this relationship in predictive models [24, 30]. Therefore, it would be reasonable to expect that spatial-temporal student networks can be useful for engineering features based on the peers a student is co-located with.

RQ3: How do network indices of spatial temporal networks relate to student performance?

2. METHODS

2.1 Datasets

Data in this study were collected from 3,915 students enrolled in five large STEM freshman courses at the University of Michigan, USA during the Fall semester of 2018. The selected courses include introductory physics, calculus, biology, chemistry, and psychology. Note that while these make up only a small fraction of all available offerings, they are considered to be foundational for a wide range of degree programs. That is, these courses serve as a gateway into the discipline, account for a significant portion of total credits registered, and are an integral part of one's academic career upon which we can leverage data collection to better understand the broad needs of incoming students and to improve instruction. The format of these courses is primarily didactic in nature, consisting of large lecture-style classes with hundreds of student enrollments. Content coverage is relatively stable between terms albeit with changing instructional teams, and the diverse student body, both in terms of demographics and measures of performance, was a key determinant in selecting these log data to represent students' first-year experience.

All data were de-identified. The dataset contained 91.7 million time-stamped entries recording log data between each device being connected to a particular WiFi access point. Each entry contained a unique user ID, a timestamp, a timestamp when a device was disconnected from a WiFi access point, a WiFi access point descriptor which (often) included a physical location such

as a building name and room number, and the device MAC address (Table 1).

Table 1. De-identified sample WiFi data

ID	Timestamp	Session End	Access point	MAC
A1234	2018-09-24 08:00:00	NA	TWC-1023	XYZ123
A1234	2018-09-24 08:02:00	NA	TWC-2013	XYZ123
B2314	2018-09-24 08:00:03	2018-09-24 08:00:55	BAHR-1210	XYZ125
C2153	2018-09-24 08:00:05	NA	CQTB-3734	XYZ121

The data were pre-processed by dropping all records generated by MAC addresses that were connected to access points within a single building, because they were likely to be stationary devices, such as computers at the libraries or lecture halls. Second, we computed a "connected time" feature for each user by subtracting two consecutive timestamps (t2-t1). For example, the connected time for user A1234 to access point TWC-1023 was 2 minutes (Table 1). The connected time feature is important for the subsequent network modeling, which requires a co-located time between any two users. Since the connected time could be biased when a device became disconnected (i.e. students left the building), we removed all data entries which contained a session end's timestamp. After the pre-processing, we retained 80.9 million records of 3,910 users, and these records were joined with the demographic information and final grades for a semester.

2.2 Analysis

2.2.1 Compute co-located time

To draw inferences about the network structure from WiFi data, we created an undirected weighted one-mode network (i.e. user-user). A tie's weight equaled to the total amount of co-located time between two users. Figure 1 visualizes the temporal changes in WiFi access points of two users on a particular date from 08:00 to 20:00. These two users spent a large amount of time in the morning at a fixed WiFi access point, possibly attending a lecture. In the afternoon, these two users shared the same access points for 2 hours. After that, each user went on about their day to different areas on campus.

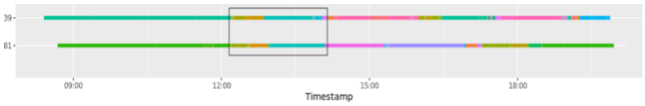


Figure 1: Temporal changes in WiFi access points of two users throughout a day. The boxed area indicates a two-hour period where these users shared the same access point

The co-located time between each pair of users was computed using the *roverlaps* package and stored in a 3910 x 3910 adjacency matrix.

2.2.2 Exponential random graph models (ERGMs)

RQ1 seeks to understand how demographic characteristics (e.g. gender, ethnicity, minority, under-representative, family income, parents' educational level), and academic performance relate to the formation of ties amongst students. Specifically, we model if students from the same background or having the same academic performance were more likely to form a connection. We used Exponential Random Graph Model (ERGM) techniques which have been used to explore homophily in network formation in educational data previously [9, 13, 14]. ERGM, also known as a p* model, is a stochastic model that specifies the probability of the entire network as a function of its network properties [31].

$$P(Y = y) = \exp(\theta' g(y)) / k(\theta)$$

- Y is the network realization;
- y is the observed network;
- g(y) is a vector of model statistics for network y;
- θ is the vector of coefficients for those statistics, and
- k(θ) represents a normalizing factor, calculated as the sum of $\exp(\theta' g(y))$ over all possible networks.

This can be expressed as the conditional log-odds of a single tie between two actors i and j:

$$\text{logit}(Y_{ij} = 1 | y_{ijc}) = \theta' \delta(y_{ij})$$

where θ is the coefficient and $\delta(y_{ij})$ is a change statistic.

To translate this into our context, ERGM was used to estimate the likelihood, expressed in conditional log-odds of two students being connected, given the similarity in their demographic characteristics and course grades. Model fit was examined with AIC and BIC (the lower the better model fit) and visual plots.

An important analytical decision was taken when weighted ties in our spatial temporal network were transformed into binary relations. To do so, we applied a filtering technique called dyadic thresholding. That is, a tie between two students would be kept when its weight was more than two standard deviations above the mean of all weights across all students. In other words, two users were considered to have a social connection when they spent a large proportion of their time on campus around each other.

To address RQ2, we applied a time window slicing technique to create separate ERGMs for a network that captured every month of activities from September to December. We then compared the changes in network homophily based on demographics and academic performance across four months.

Finally, for RQ3, network indices at the level of a node/student were incorporated in mixed-effect regression models. The models predicted grades as a function of demographics and network properties. To test the relationship between peer performance and predicted grade, we incorporated two features: 1) the average grade of the most frequently co-located peer, and 2) the average grade of five most frequently co-located peers into the model.

All the analyses were carried in R 3.6.2. ERGMs were fit using the *statnet* package [31, 32], mixed-effect regression models were run with the *lme4* package [33]. A simulated dataset and the code will be made available on Github. (https://github.com/quan3010/EDM20_Nguyen).

3. RESULTS

3.1 Network description

The data for network construction was comprised of 80.9 million log events of 3,910 users over four months. From that, we derived a weighted, undirected network with over 6.54 million weighted ties. An average co-located time between two students was 0.98 hours, with a standard deviation of 12.37 hours. This weighted graph was converted into an unweighted graph network by setting a cut-off value equal to two standard deviations about the mean, i.e. 25.74 hours. Thus, in the modelled network two users shared a tie only if they spent at least 25.74 hours together over a four-month period. The final network had 3,910 users and a total of

18,704 ties. In such a network, the median number of ties was 8 with maximum of 63 ties. For 50% of the students the range of connections was from a minimum of 3 peers to a maximum of 14. The average co-located time between two users was 120 hours, with a minimum of 25.74 hours, median of 38 hours, and a maximum of 1397.62 hours.

Table 2. Frequency statistics of demographic and grades

Gender	N	Percentage
Male	1969	50.4%
Female	1941	49.6%
Ethnicity		
White	2207	56.4%
Asian	763	19.5%
Hispanic	337	8.6%
Mixed	214	5.5%
Not Indic	197	5.0%
Black	187	4.8%
Native American	5	0.1%
Minority status		
Non-minority	2365	60.5%
Minority	1346	34.4%
International	199	5.1%
Underep stats		
Non-Underrepresented Minority	3083	78.8%
Underrepresented Minority	628	16.1%
International	199	5.1%
Family income		
> $200,000	1043	26.7%
$150,000-$199,999	355	9.1%
$100,000-149,999	563	14.4%
$75,000-$99,999	243	6.2%
$50,000-$74,999	266	6.8%
$25,000-$49,999	366	9.4%
< $25,000	217	5.6%
NA	847	21.7%
Grade_letter		
A-, A, A+	1295	33.1%
B-, B, B+	1671	42.7%
C-, C, C+	664	17.0%
Below D	140	3.6%
Withdraw	120	3.1%
NA	20	0.5%

Table 2 provides descriptive statistics for 3,910 students in this study. There was a rough balance in the number of female and male students. This is important since homophily can occur at random, for instance when a relative size of a subgroup is markedly different. White was the most frequent ethnicity, followed by Asian and Hispanic. A third of the sample identifies as an ethnic minority and 16.1% was categorized as under-represented minority. The family income distribute are right-skewed with over a quarter of students report household income of over $200,000[1]. Academic performance in this semester followed a bimodal distribution with of the majority of students performed at the A-range and B-range.

3.2 Homophily based on demographics and grades

Table 3 reports the results of three ERGM models. Model 1 serves as the baseline model, which accounts for the density of the network. The log-odds of a tie was -6.01 which translates to a probability of a tie exists equal to 0.24% (i.e. 18,704 ties divided by a total of 7.64 million possible ties).

In model 2, we added five nodal attributes, including gender, ethnicity, ethnic minority status, under-represented minority status, and family income, to explore homophily related to demographics. Our results showed that students from the same gender were more likely to form a tie than those with different gender, with the probability of a same-gender tie being 62%. Ethnicity and underrepresented minority status of the student did not have any statistically significant effect on the formation of network ties. This may be explained by the effect of the minority variable, which already accounted for ethnicity and unrepresented groups. Although a social connection was more likely to exist between students from the same minority group (i.e. non-minority, international, minority), the effect was marginal with a probability of only 54%. Family income also had a small effect on the formation of ties. The probability of ties to exist between two users with the same family income was 53%.

Table 3. Homophily effects of demographics and grades

	Model 1	Model 2	Model 3
ties	-6.010***	-6.375***	-6.444***
	(0.007)	(0.016)	(0.017)
gender		**0.479*** **	**0.479*** **
		(0.015)	(0.015)
ethnic		-0.006	-0.008
		(0.022)	(0.022)
minority		**0.157*** **	**0.161*** **
		(0.021)	(0.021)
underrep		0.004	-0.004
		(0.018)	(0.018)
family_income		**0.135*** **	**0.132*** **
		(0.021)	(0.021)
grade_letter			**0.211*** **
			(0.015)
AIC	262,286	261,101	260,913
BIC	262,300	261,184	261,010

[1] Household income is self-reported on admissions data.

Note: *p<0.1; **p<0.05; ***p<0.01
Coefficients calculated in log-odds, standard errors in brackets

Finally, in model 3, we added student's grades to examine homophily related to academic performance. The probability of a tie among same-grade students was 55%. To conclude, we observed a strong homophily network effect in gender, and marginal effects in minority identity, family income, and academic performance. Spatial-temporal networks also captured the commonly observed patterns of social homophily. This suggests that spatial-temporal networks reflected the social connections underpinning the co-location patterns.

The measures of homophily based on demographics have important implications to the understanding of diversity and inclusivity in higher education. The mere presence of structural diversity in student body (i.e. the proportional representation of groups of students from different backgrounds) does not guarantee the interactions between these diverse groups (Puritty et al., 2017). Homophily measures could serve as an indicator of how diverse and inclusive the social interactions between students are. A highly homophilous network could signal social segregation, and to some extent, inequality in student body as students are less likely to form a connection with peers who are demographically different than themselves. The use of WiFi data could support the design of physical spaces/educational activities that increase the likelihood of spatial co-occurrences between diverse groups of students.

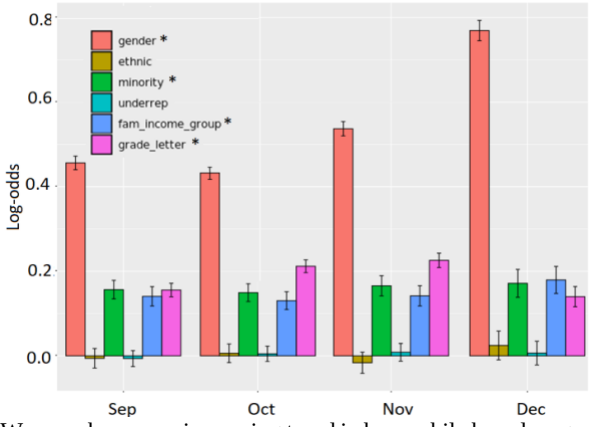
However, we are careful to draw inferences as to what the homophily represents. Our models do not control for types of building, or events that take place on campus. It is plausible that spatial temporal networks capture both the networks formed based on foci of activity (classes, living arrangements, cafeteria visits for students with similar schedules) as well as social ties. For instance, gender homophily could be explained by the majority of freshman students sharing their living space with same-gender peers in a residential building on campus. In this case, co-located time between roommates and dormmates would be the highest among freshmen. We did not find any evidence of homophily between different ethnicities per se. However, we observed homophily between different ethnic identities, such as ethnic minority (i.e. Black, Asian, Mixed, Hispanic), ethnic non-minority (i.e. White), and international (i.e. mostly Asian). In other words, there was evidence for inter-ethnic co-location within ethnic minorities.

As can be seen in the model, the addition of the terms decreased the AIC/BIC suggesting improved model fit. We did not manage to fit any of the conventional closure terms, such as popularity (e.g. geometrically weighted degree distribution) or transitivity (e.g. geometrically weighted edgewise shared partners), into the model. Visual examination of the goodness of fit suggested that the model was fit in predicting dyadic-level observations but was limited in reproducing the network structure. These results suggest that the model either requires to add control variables about the events/reasons for co-location (e.g. lectures, Thanksgiving breaks, exam periods), or that the networks need to be separated to have a more elaborate operationalization of co-location (e.g. residential building, libraries, classrooms).

3.3 Temporal changes in social networks

To examine the changes in the homophily over time, we ran the ERGM model with the same specification for a network capturing co-location in each month (Sep, Oct, Nov, Dec). The coefficients of each model were visualized in Figure 2.

Figure 2. Temporal changes in homophily effects of demographics and grades on network formation (* p < 0.01)



We can observe an increasing trend in homophily based on gender over time. The probability of a same-gender tie increased from 61% in September to 69% in December. There was a small increase in the homophily based on grade in October and November but it then decreased in December. The homophily effect of minority identity and family income remained constant over time.

One potential explanation for the increasing trend in gender-based homophily is that students started expanding their social circle with people in the same dorm hall/residential building, who are likely to have the same gender. This could also be explained by the participation in fraternity and sorority activities for freshman. As a result, we observed an increase in same-gender co-location over time as students formed new connections within a fraternity and sorority. Finally, previous studies [29, 30] also observed the intersectional nature of grade-based performance, i.e. high-performing boys are likely to form ties with high performing boys, and the same applies to girls. The consistent trend in performance-based homophily could be explained by the fact that this is the first semester and not only are students new to the institution, but university academic performance was generally not available until the end of the semester. Results suggest that it could be interesting to examine the temporal changes in performance-based homophily over a longer time period, especially in sophomore and senior students.

This finding has important implication to the research of social interactions between students. More often than not, social relations in educational research are collapsed under a static and dichotomous category (e.g. yes/no). In reality, the formation of social relations is a highly dynamic and time-variant process. For example, students could become closer with certain peers while more distant with others as time goes by. Students' social circle could be more elastic during their freshman year but gradually form a close-knit group as they approach their senior year. The networks inferred from WiFi data allow us to explores many questions about the evolution in social interactions between students over time, which could not previously be answered with self-report social network surveys.

3.4 Predicting academic performance

We applied mixed-effect regression models to control for the heterogeneity between courses (Table 4). Grade letters were converted into numeric format as per institutional guidelines, with a maximum value of 4.0. Our findings indicated that in the courses we studied, male students on average achieved 0.08 grade

points higher than female students. Compared to students with a family annual income over $200,000, which accounted for a quarter of our dataset, students with a family income of $75,000, $50,000, and $25,000 had 0.13, 0.30, and 0.43 grade points lower respectively. The effect of family income became marginal and non-statistically significant once it is above $100,000. Students from an under-represented minority (i.e. Black, Hispanic, and Native American) also had on average 0.30 grade points lower than a non-underrepresented minority (i.e. Asian, Mixed, and White).

All three network indices had a positive and statistically significant relation with academic performance. On average, each additional tie increased a student's final course grade by 0.014 grade points. For each grade point increase in the most frequently co-located peer, the student's grade increased by 0.07 grade points. For each additional point increase in the average grade of the five most frequently co-located peers, a student's grade increased by 0.15 grade points. It is important to note that our results so not imply a causal relationship. The finding could be explained by a homophily effect (i.e. students co-located with similarly performed peers) or a roommate effect (i.e. performances of co-located peers influence a student's performance).

Table 4. Effects of demographics and network on grades

	Model 1	Model 2	Model 3	Model 4	Model 5
Male	0.055*	0.071**	0.075**	0.082**	0.080**
	(0.023)	(0.025)	(0.025)	(0.025)	(0.025)
Ethnicity (ref= White)					
Mixed	-0.036	-0.111	-0.065	-0.060	-0.070
	(0.135)	(0.177)	(0.175)	(0.174)	(0.174)
Asian	0.033	0.026	0.070	0.076	0.059
	(0.122)	(0.166)	(0.164)	(0.163)	(0.163)
Black	-0.358*	-0.304	-0.269	-0.254	-0.253
	(0.149)	(0.189)	(0.187)	(0.187)	(0.187)
Hispanic	-0.041	-0.049	-0.025	-0.012	-0.019
	(0.143)	(0.183)	(0.181)	(0.181)	(0.180)
Native Am	-0.295	-0.239	-0.141	-0.143	-0.129
	(0.357)	(0.374)	(0.369)	(0.368)	(0.367)
Ref = Non-minority					
Internatnl	0.239*	0.174	0.173	0.174	0.183
	(0.109)	(0.151)	(0.149)	(0.149)	(0.149)
Minority	0.043	0.067	0.028	0.020	0.033
	(0.125)	(0.168)	(0.166)	(0.165)	(0.165)
Ref = Non-underrepresented minority					
Underrep	-0.319***	-0.297***	-0.314***	-0.305***	-0.304***
	(0.081)	(0.090)	(0.089)	(0.089)	(0.089)
Family income (ref=above $200,000)					
$199,999		0.007	-0.007	-0.012	-0.010
		(0.042)	(0.042)	(0.042)	(0.042)
$149,999		-0.082*	-0.091*	-0.085*	-0.082*
		(0.036)	(0.035)	(0.036)	(0.036)
$99,999		-0.086	-0.081	-0.094*	-0.089
		(0.048)	(0.047)	(0.048)	(0.047)
$74,999		-0.133**	-0.139**	-0.131**	-0.131**
		(0.048)	(0.047)	(0.047)	(0.047)
$49,999		-0.280***	-0.309***	-0.303***	-0.298***
		(0.042)	(0.042)	(0.042)	(0.042)
$25,000		-0.460***	-0.452***	-0.446***	-0.429***
		(0.053)	(0.053)	(0.053)	(0.053)
No. of ties			**0.014*** **	**0.014*** **	**0.014*** **
			(0.002)	(0.002)	(0.002)
Closest peer				**0.106*** **	**0.072*** **
				(0.016)	(0.019)
5 close peers					**0.150*** **
					(0.040)
Constant	3.114***	3.193***	3.059***	2.717***	2.345***
	(0.130)	(0.129)	(0.142)	(0.149)	(0.176)
Obs	4,422	3,554	3,554	3,500	3,500
AIC	9,872.995	7,945.483	7,871.099	7,733.857	7,726.519
BIC.	9,956.122	8,068.999	8,000.792	7,869.388	7,868.211
Note:	*p<0.05; **p<0.01; ***p<0.001				

4. CONCLUSION

This paper explores the use of WiFi data of 3,910 students in Fall 2018 in understanding student physical on-campus connections. Specifically, we explore if spatial-temporal student networks reflect homophily based on demographics and academic performance expected in social networks. Network connections were inferred when two users exhibited a high level of co-located time (i.e. connecting to the same WiFi access point in the same time window). We found evidence of homophily with regards to gender, ethnic minority identity, family income, and academic performance. Gender-based homophily is particularly interesting, given that the composition of the student body has equal share of both genders and that this homophily increased significantly over time. This suggests that observed homophily is not baseline, but largely structural. That is, the organization of physical space, as well as curricular and extracurricular activities may create opportunities for gender-based homophily on campus. Exploring this further may be useful in understanding the effect of various institutional (e.g. gender-based meetups, structured study sessions, or mentoring workshops) and non-institutional (e.g. gender-based social activities, such as fraternity and sorority functions and enrollments) activities on the development of friend and support networks.

In addition, we found that the number of ties and the average performance of the most frequently co-located peer(s) were predictive of academic performance. This is in line with extant literature on self-reported peer effects [29], or the effects of peers observed from academic records [24]. Contextualizing this relationship and determining signals for specific causal activities is a clear next step.

From a theoretical perspective, our results confirmed homophily with regards to demographics and academic performance. At the same time, we extended the findings to capture the temporal changes in homophily within a semester. Our findings suggest that the tendency to form (co-located) connections may vary over time and more longitudinal studies are needed to understand the mechanism behind dynamic homophily.

From a methodological perspective, we demonstrated a novel application of spatial-temporal data in the study of student social networks, which have primarily relied on self-reports and log-data from discussion forums. This opens up a new venue to capture social interactions between students on campus on a large

scale and with fine-grained granularity. Importantly, this can be achieved without the need to collect additional data beyond what has been already collected by the university wireless networks. Location data inferred from WiFi access points can be considered as less invasive than using mobile phone's GPS or location-sensors to track users' location [15, 16]. Future research could combine self-report, discussion forum data, and location-based data to form a more holistic picture of student social networks and to triangulate findings from multiple data sources.

From a practical perspective, this study highlighted several factors that determine the formation of network among college students as well as their effect on academic performance. Such results may be useful to institutions in designing or evaluating location-based initiatives to promote gender, ethnicity, and culture diversity and inclusivity on campus, as well as to support ethnic minority, underrepresented minority in social integration during their time in college. There are opportunities to better understand the impacts of learning communities (e.g. themed residences for groups of students, such as Women in STEM communities), of co-curricular activities and their placement on campus (e.g. guest speakers or academic support groups), and even architectural planning (e.g. the relationship between dormitories and classrooms or libraries) through these methods.

4.1 Limitations

The data used does note capture the use of non-university run network (e.g. cellular networks), when students choose to go offline (e.g. intentionally by powering down their phone or due to low battery), or in spaces on campus without access to university network. There is also an inherent messiness which comes with the use of multiple or shared devices, the former of which is very common and increasing with the use of wearables. Network inference based on co-located time is further be biased when students co-locate by random chance or by sharing common activities (i.e. attending lectures, going to the libraries, going to the gym) but do not interact with one another. Similarly, it is possible for students to be in completely different rooms yet connected to the same access point depending upon the wireless network and building topologies, introducing further noise to social network models. As a result, there might be hidden bias when using networks inferred from location data for predictive purposes. More sophisticated network inference techniques may be helpful in understanding this, such as weight/tie reshufflings or spatial/temporal simulations [34], and better cataloging of network endpoints (e.g. classroom, office, hallway) may be helpful in modeling social network relationships.

Finally, the modeling techniques used with the limited dataset chosen required significant computing power. More fine-grained temporal analyses (e.g. weekly or daily models), a longer time frame (e.g. a full academic year or throughout the students' academic career), and increased data (e.g. from more courses and non-freshman students) will only increase the need for computational power.

4.2 Concerns with the Use of Wi-Fi Data

WiFi data is highly sensitive data and the security of the collection, storage, and analysis of such data is of utmost importance. As is appropriate, we sought IRB oversight of our use of this data and worked with institutional data governance teams to ensure the data we received was appropriately stored, was de-identified, and was as minimal as possible to support our analyses.

At the same time, we feel it incumbent upon us to note that research access to such data is under threat by the potential misuse of educational location data for non-research purposes, which does not have to undergo IRB review. Specifically, some have begun to incorporate location data into formative evaluation of students. Location data is captured not only through WiFi, but also Bluetooth beacons and student mobile application software (which may be required), and has been used in an identifiable way to assign students grades based on location (attendance in class) [35]. While there are broad discussions of agency, privacy, paternalism, and ethics which the authors have positions on, the purpose of this section of the paper is to raise the importance such data has in understanding teaching and learning, and to encourage researchers in the field of Educational Data Mining (EDM) to voice opinions on the value of de-identified location data and it's use in educational research.

5. ACKNOWLEDGMENTS

This work was funded under the Holistic Modeling of Education (HOME) project funded by the Michigan Institute for Data Science (MIDAS). We would also like to thank the UM's ITS team for their support with data access.

6. REFERENCES

[1] McPherson, M., Smith-Lovin, L. and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27, 1, 415-444.

[2] Grunspan, D. Z., Wiggins, B. L. and Goodreau, S. M. 2014. Understanding classrooms through social network analysis: A primer for social network analysis in education research. *CBE—Life Sciences Education*, 13, 2, 167-178.

[3] Xu, Y., Gitinabard, N., Lynch, C. and Barnes, T. 2019. What You Say is Relevant to How You Make Friends: Measuring the Effect of Content on Social Connection. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*. Montreal, Canada

[4] Xu, Y., Lynch, C. F. and Barnes, T. 2018. How Many Friends Can You Make in a Week?: Evolving Social Relationships in MOOCs over Time. In *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)*. Buffalo, New York

[5] Wise, A. F. and Cui, Y. 2018. Learning communities in the crowd: Characteristics of content related interactions and social relationships in MOOC discussion forums. *Computers & Education*, 122, 221-242.

[6] Fincham, E., Gašević, D. and Pardo, A. 2018. From Social Ties to Network Processes: Do Tie Definitions Matter? *Journal of Learning Analytics*, 5, 2, 9–28-29–28.

[7] Dowell, N. M., Skrypnyk, O., Joksimovic, S., Graesser, A. C., Dawson, S., Gaševic, D., Hennis, T. A., de Vries, P. and Kovanovic, V. 2015. Modeling Learners' Social Centrality and Performance through Language and Discourse. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*. Mandrid, Spain

[8] Rabbany, R., Elatia, S., Takaffoli, M. and Zaïane, O. R. 2014.Collaborative Learning of Students in Online Discussion Forums: A Social Network Analysis Perspective. Springer International Publishing.

[9] Zhu, M., Bergner, Y., Zhang, Y., Baker, R., Wang, Y. and Paquette, L. 2016. Longitudinal engagement, performance, and social connectivity: a MOOC case study using exponential random graph models. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*. 223–230. Edinburgh, United Kingdom

[10] Rienties, B. and Tempelaar, D. 2018. Turning Groups Inside Out: A Social Network Perspective. *Journal of the Learning Sciences*, 27, 4, 550-579.

[11] Chen, B., Chang, Y.-H., Ouyang, F. and Zhou, W. 2018. Fostering student engagement in online discussion through social learning analytics. *The Internet and Higher Education*, 37, 21-30.

[12] Poquet, O. and Dawson, S. 2016. Untangling MOOC learner networks. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*. 208-212.

[13] Poquet, O., Dowell, N., Brooks, C. and Dawson, S. 2018. Are MOOC forums changing? In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. 340-349.

[14] Joksimović, S., Manataki, A., Gašević, D., Dawson, S., Kovanović, V. and De Kereki, I. F. 2016. Translating network position into performance: importance of centrality in different network configurations. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*. 314-323.

[15] Stopczynski, A., Pentland, A. S. and Lehmann, S. 2018. How Physical Proximity Shapes Complex Social Networks. *Scientific Reports*, 8, 1, 17722.

[16] Eagle, N., Pentland, A. S. and Lazer, D. 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106, 36, 15274-15278.

[17] Sevtsuk, A. 2009.Mapping the MIT campus in real time using WiFi. IGI Global.

[18] Zhou, M., Ma, M., Zhang, Y., SuiA, K., Pei, D. and Moscibroda, T. 2016. EDUM: classroom education measurements via large-scale WiFi networks. In *Proceedings of the 2016 acm international joint conference on pervasive and ubiquitous computing*. 316-327.

[19] Sarkar, S., Carpenter, B., Bader-El-Den, M. and Knight, A. 2016. Where students go and how they do: Wi-Fi location data versus academic performance. In *Proceedings of the 9th International Conference on Human System Interactions (HSI)*. 45-51.

[20] Hang, M., Pytlarz, I. and Neville, J. 2018. Exploring student check-in behavior for improved point-of-interest prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 321-330.

[21] Smirnov, I. and Thurner, S. 2017. Formation of homophily in academic performance: Students change their friends rather than performance. *PloS one*, 12, 8.

[22] Stehlé, J., Charbonnier, F., Picard, T., Cattuto, C. and Barrat, A. 2013. Gender homophily from spatial behavior in a primary school: A sociometric study. *Social Networks*, 35, 4, 604-613.

[23] Smith, J. A., McPherson, M. and Smith-Lovin, L. 2014. Social distance in the United States: Sex, race, religion, age, and education homophily among confidants, 1985 to 2004. *American Sociological Review*, 79, 3, 432-456.

[24] Gardner, J. and Brooks, C. 2018. Coenrollment networks and their relationship to grades in undergraduate education. In *Proceedings of the 8th international conference on learning analytics and knowledge*. 295-304.

[25] Gitinabard, N., Khoshnevisan, F., Lynch, C. F. and Wang, E. Y. 2018. Your actions or your associates? Predicting certification and dropout in MOOCs with behavioral and social features. In *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)*. Buffalo, New York

[26] Fire, M., Katz, G., Elovici, Y., Shapira, B. and Rokach, L. 2012. Predicting Student Exam's Scores by Analyzing Social Network Data. In *Proceedings of the International Conference on Active Media Technology* 584-595. Berlin, Heidelberg

[27] Vaquero, L. M. and Cebrian, M. 2013. The rich club phenomenon in the classroom. *Scientific Reports*, 3, 1, 1174.

[28] Kretschmer, D., Leszczensky, L. and Pink, S. 2018. Selection and influence processes in academic achievement—More pronounced for girls? *Social Networks*, 52, 251-260.

[29] Sacerdote, B. 2001. Peer effects with random assignment: Results for Dartmouth roommates. *The Quarterly journal of economics*, 116, 2, 681-704.

[30] Gašević, D., Zouaq, A. and Janzen, R. 2013. "Choose your classmates, your GPA is at stake!" The association of cross-class social ties and academic performance. *American Behavioral Scientist*, 57, 10, 1460-1479.

[31] Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M. and Morris, M. 2008. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of statistical software*, 24, 3, nihpa54860.

[32] Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M. and Morris, M. 2008. statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of statistical software*, 24, 1, 1548.

[33] Bates, D., Mächler, M., Bolker, B. and Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67, 1, 1-48.

[34] Poquet, S., Tupikina, L. and Santolini, M. 2019. Are forum networks social networks? A methodological perspective. In *Proceedings of the 10th International Conference of Learning Analytics & Knowledge (LAK20)*. in press. Frankfurt, Germany

[35] Harwell, D. 2019. *The Washington Post*. Retrieved from https://www.washingtonpost.com/technology/2019/12/24/colleges-are-turning-students-phones-into-surveillance-machines-tracking-locations-hundreds-thousands/