Lance Tan
5/8/19
SOCY 167 Final Paper

Homophily in academic majors and roommate relationships: a network analysis of Yale student directory data

Introduction

Choosing an academic major of study is one of the most important milestones in the career of a university undergraduate. Americans students most frequently cite improved employment prospects as the reason they are in college (Rampell 2015), and a student's choice of academic major strongly directs their later career path. Consequently, choosing an academic major is a significant source of stress to college undergraduates. An author of a recent Yale Daily News opinion (Jin 2018) notes that "on campus, declaring your major is considered just that: major"; she writes that "to choose 'incorrectly' was to surrender the privilege of higher education." This sentiment is echoed in student publications at several other schools (Nielsen 2016) (Daly 2013). Clearly, choosing a major is a momentous decision in a student's life.

But how do university students actually choose their majors? Many authorities, including the College Board (Anon 2018), suggest that students consider academic and personal interests, personality traits, career goals, and the advice of academic advisors, hinting at an extremely individual and introspective process. However, I propose that there is a significant social aspect to the process of choosing an academic major: a student's choice of major both influences and is influenced by their peers' choices, and the process is intimately related to the social network of friends that the student is embedded in. The same Yale Daily News opinion (Jin 2018) articulates how "to be undeclared is to be in a state of discomfort[,] but to be undeclared in a sea of passionate, driven, decided people — is painful": the stress due to not having declared a major is

at least partially due to having many friends who have already declared. To explore this theory, I use data from Yale University's undergraduate student directory to conduct a social network analysis of how Yale students' majors compare to their roommates' majors; since Yale sophomores, juniors, and seniors choose their own roommates, I use roommate relationships as a proxy for friendships. I find evidence that as Yale students approach graduation, they associate more strongly with peers majoring in similar disciplines; and that within the Yale sophomore class, students in certain majors tend to room with students in similar majors to a statistically significant ($p < .01$) extent.

Hypothesis and Data

In social network theory, homophily is the principle that if two social actors (individuals or collectives) are connected, then they are more likely to have common characteristics, and vice versa. In other words, two individuals closely connected within a social network are likely to have the same attributes or norms: the idea that, as Kadushin puts it, "'birds of a feather flock together'". (Kadushin 2012:18–20) Accordingly, I hypothesize that on average, students that live together will tend to major in similar disciplines. Kadushin further differentiates between two kinds of homophily. In one, common values or interests may bring people together, or the reverse, friends might adopt each other's interests and beliefs; in the second, presence in the same physical or structural location (such as a classroom or business) may cause common attributes, or the reverse. This study, about how students' choice of major compare across roommate relationships, is most similar to the first type. Kadushin also emphasizes the ongoing nature of homophily, noting that "homophily is a process" and "the feedback between network structure and individual preference thus becomes especially noticeable over time." With this in

mind, I further hypothesize that evidence of homophily in academic major will be most apparent in the senior class.

The data for this study was collected from "Yale Face Book" an online, internal directory of Yale undergraduates.[1] The directory includes name and contact information for all students by default; a student may remove any of their information from the directory if they choose (but may not remove their entry from the directory entirely). An ad-hoc script was used to access all entries in the directory (n=6139) and parse each student's name, year of graduation, declared major, and on-campus room number (or off-campus address). All Yale undergraduate students living on campus live either in single rooms, or in suites with about five and up to ten other students of the same class. This study focuses on these same-suite relationships, and will call two students "roommates" if they live in the same suite. Directory entries missing information, or for students living off campus or in single rooms on campus, were eliminated, since the directory does not contain roommate information for off-campus residences. The remaining students were grouped by suite, forming a dataset for this study (n=4174).

Table 1: Student directory information and extracted dataset

Class (graduation year)	Number in directory	Number missing information/live off campus	Number in single rooms	Number in dataset
2022	1597	24	22	1551
2021	1629	113	81	1435
2020	1426	476	267	683
2019	1397	677	215	505
Unknown	90	90		
Total	6139	1380	585	4174

[1] https://students.yale.edu/facebook/

Due to new residential spaces opening in 2017, the sophomore and first-year classes are larger than the junior and senior classes. Furthermore, a higher proportion of Yale juniors and seniors live off campus or live in single rooms. Due to both of these factors, juniors and seniors are underrepresented in this dataset, which will limit conclusions that can be drawn from this data.

Methods and Analysis

For each of the first-year through senior classes, a bipartite graph was constructed. Each node in the graph represents either a suite of students or an academic major (including "undeclared" for students without a declared major). The weight of an undirected edge (s, m) between a suite s and a major m represents the number of students in the suite s who declared m as their major. Since this graph has two types of nodes, and its edges only connect nodes of different type, the graph is bipartite. Following (Beckfield 2010), these bipartite graphs were then collapsed into additional unimodal graphs, where each node is a major, and the weight of an edge (m_1, m_2) represents the number of roommate relationships in the class between students studying m_1 and students studying m_2. In other words, the weight of (m_1, m_2) is the sum for all students s studying m1 of the number of s's roommates studying m_2. This collapsed graph has loops (edges that connect a vertex to itself), representing the number of roommate relationships where both students are majoring in the same field.

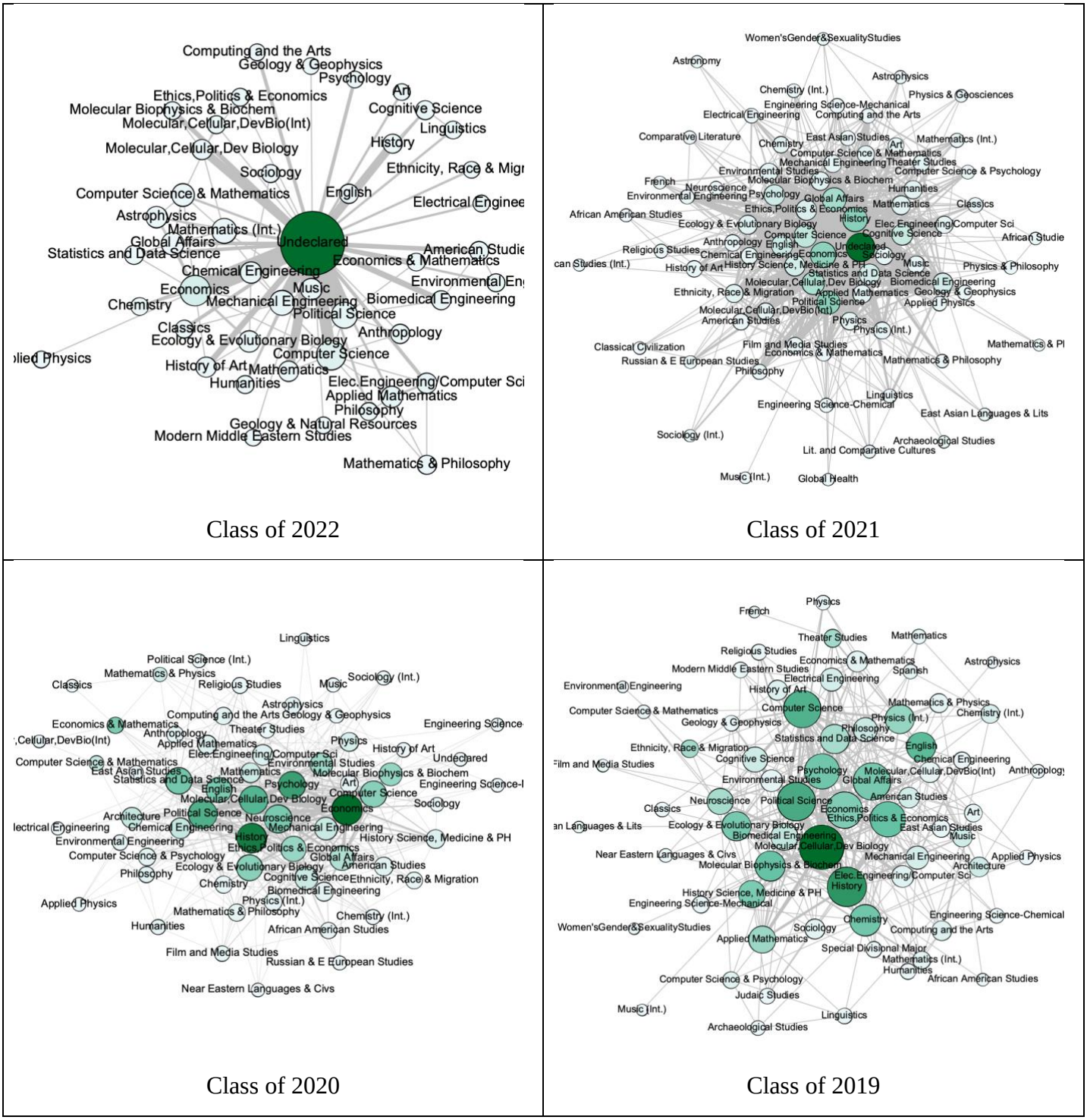
Part 1: Visualization, density and centrality analysis of roommate-major networks

The four collapsed graphed were visualized using the Yifan Hu layout (Hu 2004). This algorithm models the graph drawing as a system of springs between neighboring nodes, then seeks to minimize the energy of the system. As a result, nodes with more neighbors—majors linked via roommate relations to more other majors—are pushed to the center of the graph, which improves visualization.

In the visualization, larger nodes represent majors with more students, and darker-colored nodes represent majors with higher betweenness centrality. There exists a shortest path between any two nodes in a connected graph; the betweenness centrality of a node is the number of these shortest paths that pass through the node. There are several other measures of node-level centrality, including degree, closeness, and eigenvector centrality; here, betweenness centrality was chosen since in this context, a shortest path has a more intuitive interpretation: the length of a shortest path is the number of hops needed to move from one major to another via roommate relationships.

To compare the tendency of roommates to study similar majors between the four classes, I calculate the density of the bipartite network, density of the unimodal network, as well as average path length in the unimodal network. To calculate density of a bipartite graph, I follow (Beckfield 2010), who notes that in a bipartite graph with l and r nodes, edges are only possible between nodes of different type, so the correct denominator for is lr.

Figure 1: Unimodal collapsed graph for each class



Loops in graphs are omitted.

Table 2: Density and average path length

Class	Bipartite graph (suites and majors)	Collapsed graph (suites only)		
	Density	Density	Average path length	n
2022	0.0322	0.101	1.961	45
2021	0.0502	0.267	1.759	74
2020	0.0508	0.238	1.873	62
2019	0.0398	0.155	2.143	67

Since a large number of students of the class of 2022 have not declared a major, it does not reflect the trend exhibited by the other three classes. While the other three collapsed graphs appear well-connected, with no single node dominating the center of the graph, the collapsed graph for the class of 2022 is very star-shaped, with almost all of the edges connecting an undeclared student with a student with a declared major, rather than two students both with declared majors. This is reflected in the low graph densities and average collapsed graph path length close to 2, where many shortest paths are the "undeclared" node connecting two other majors.

The sophomore through senior classes show evidence of homophily in Yale roommates' majors: that Yale students tend to associate with peers majoring in similar disciplines. The graphs from the sophomore to senior classes become progressively less dense, and the average path length increases, implying that as students near graduation, they increasingly prefer to room with peers in a smaller subset of other majors. Indeed, this agrees with the principle that evidence of homophily becomes more evident over time: undergraduate seniors have had the most time to make and form college friendships. Concordantly, except for the first-year class, the senior class has the lowest graph density for both the bipartite and collapsed graphs, and the senior class collapsed graph has the largest average path length. Student directory data thus exhibits network effects: who a college student lives with does depends on their roommate's major.

However, a confounding factor is that fewer juniors and seniors are included in this analysis than sophomores. Whether the networks for the junior and senior class are also less dense because a smaller number of students captures fewer varieties of roommate relationships is unclear.

Part 2: Statistical analysis of roommate relationships within sophomore class

In light of this, I more closely analyze the sophomore class to determine whether there is stronger evidence of homophily in the academic majors of Yale roommates. The sophomore class was chosen since there is more data for the sophomore class than the junior of senior classes; the first-year class was excluded because of the large number of undeclared first-year students.

I use a chi-squared test to determine whether the distribution of the roommates of students in a particular set of majors is different from the roommates of Yale sophomores as a whole. The chi-squared goodness-of-fit statistic for categorical data is:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where k is the number of categories, O_i is the number of observations of type i, and E_i is the expected number of observations of type i. If students majoring in m chose roommates without regards to academic major, then we would expect the distribution of roommates of students studying m to be the same as the distribution of roommates of all Yale students, scaled by the number of students studying m. In this case, the statistic follows a χ^2 statistical distribution with degree of freedom equal to $k - 1$; large values of the statistic indicate more significant differences from the no-relationship case.

For this analysis, I sort all Yale majors into six groups: physical sciences, engineering, and mathematics; arts, philosophy, and humanities; life and environmental sciences; area, ethnic and multidisciplinary studies; social sciences and economics; and undeclared students. There are two reasons for this. First, there are majors that appear in the student directory under multiple names, for example, intensive tracks within majors such as sociology. Second, due to the E_i term in the denominator, the statistic will be less accurate if there are categories where a small number of students is expected (due to the major being less popular). This categorization was based on categorizations of fields of study from several sources[2] and is described in an appendix.

Table 2: Residuals $(O_i - E_i)$ and chi squared statistic for each category.

	Physical Science/ Engineering/Math	Art/Philosophy/ Humanities/History	Life/Environmental Sciences	Area/Ethnic/ Multidisciplinary	Social Sciences/ Economics/Political Science	Undeclared	X^2	p
Physical Science/ Engineering/Math	40.92	-15.10	2.66	-22.42	-0.32	-5.74	16.40	0.006
Art/Philosophy/ Humanities/History	-15.10	21.41	2.21	7.67	5.89	-22.08	7.47	0.187
Life/Environmental Sciences	2.66	2.21	3.68	4.83	-8.37	-5.01	1.55	0.908
Area/Ethnic/ Multidisciplinary	-22.42	7.67	4.83	12.40	13.97	-16.47	22.80	< 0.001
Social Sciences/ Economics/Political Science	-0.32	5.89	-8.37	13.97	17.78	-28.96	7.26	0.202
Undeclared	-5.74	-22.08	-5.01	-16.47	-28.96	78.26	21.28	< 0.001

[2] http://www.act.org/content/act/en/research/reports/act-publications/college-choice-report-class-of-2013/college-majors-and-occupational-choices/college-majors-and-occupational-choices.html; https://bigfuture.collegeboard.org/majors-careers

The table of residuals shows that the number of students with roommates majoring in the same category is much higher than would be expected were there no relationship, for almost all of the categories. These differences are statistically significant ($p < .01$) for three of the six categories, strongly suggesting that homophily does occur within Yale student social networks, in that students tend to choose roommates majoring in similar fields as themselves. In particular, undeclared Yale sophomores are extremely likely to live with other undeclared sophomores. This aligns with the sentiment expressed in the Yale Daily News opinion (Jin 2018) that much of the discomfort associated with being an undeclared student is due to social pressure, either actual or perceived, from peers.

To test the extent to which these results are caused by this specific categorization of majors, the analysis was repeated with a different number of categories, and similar results were obtained. However, improvement to this method remains possible in that the categories of majors used may not reflect the social processes that actually occur in students' social networks.

Conclusions

Data from a Yale University undergraduate student directory was obtained and analyzed through a social network lens, showing that undergraduate students tend to choose roommates that are studying the same or similar major as themselves. This instance of homophily appears both within fields of study and across different graduating classes. Furthermore, this tendency for students to associate along academic interests appears to increase as graduation nears—in agreement with the principle that since homophily is an ongoing and self-reinforcing process, evidence of it strengthens with time. While choosing an academic major—like any other

significant life milestone—is often conceived as an individual process, these findings suggest that the choice is also a social process, deeply connected with the student's social network.

This study also demonstrates that with sensitivity to the underlying context, it is possible for data from contact information directories to reveal useful and significant social network findings. This approach may be a viable alternative or complement to other sources of data, such as surveys. In this sense, social network principles and techniques apply to a wide, even surprising, range of facets of human life.

Bibliography

Anon. 2018. "The Ultimate Guide to Choosing a Major." *The College Board Blog*. Retrieved May 7, 2019 (https://blog.collegeboard.org/the-ultimate-guide-to-choosing-a-major).

Beckfield, Jason. 2010. "The Social Structure of the World Polity." *American Journal of Sociology* 115(4):1018–68.

Daly, Christina. 2013. "The Stress of Picking a Major (And Why It's Not That Big of a Deal) – Dukes Declassified." Retrieved May 7, 2019 (https://sites.jmu.edu/molloyfall13/the-stress-of-picking-a-major-and-why-its-not-that-big-of-a-deal/).

Hu, Yifan. 2004. "Efficient and High Quality Force-Directed Graph Drawing."

Jin, Grace. 2018. "Life, Undeclared." *Yale Daily News*, December 4.

Kadushin, Charles. 2012. *Understanding Social Networks: Theories, Concepts, and Findings*. New York: Oxford University Press.

Nielsen, Alyssa. 2016. "Choosing a Major Can Be Major Stress for Students." *The Daily Universe*. Retrieved May 7, 2019 (https://universe.byu.edu/2016/03/28/choosing-major-can-be-major-stress-for-students1/).

Rampell, Catherine. 2015. "Why Do Americans Go to College? First and Foremost, They Want Better Jobs." *Washington Post*, February 17.

Appendix: Categorization of majors for Part 2 of analysis

Category 1: Physical Science, Engineering, Mathematics

'Applied Mathematics', 'Applied Physics', 'Astronomy', 'Astrophysics', 'Biomedical Engineering', 'Chemical Engineering', 'Chemistry', 'Chemistry (Int.)', 'Computer Science', 'Computer Science & Mathematics', 'Computer Science & Psychology', 'Elec.Engineering/Computer Sci', 'Electrical Engineering', 'Engineering Science-Chemical', 'Engineering Science-Electrical', 'Engineering Science-Mechanical', 'Geology & Geophysics', 'Mathematics', 'Mathematics & Physics', 'Mathematics (Int.)', 'Mechanical Engineering', 'Physics', 'Physics & Philosophy', 'Physics (Int.)', 'Statistics and Data Science'

Category 2: Arts, Philosophy, Humanities, History

'Architecture', 'Art', 'Classical Civilization', 'Classics', 'Computing and the Arts', 'English', 'Film and Media Studies', 'History', 'History Science, Medicine & PH', 'History of Art', 'Humanities', 'Mathematics & Philosophy',
'Music', 'Music (Int.)', 'Philosophy', 'Theater Studies'

Category 3: Area, Ethnic and Multidisciplinary Studies

'African American Studies', 'African Studies', 'American Studies', 'American Studies (Int.)', 'Comparative Literature', 'East Asian Languages & Lits', 'East Asian Studies', 'Ethics,Politics & Economics', 'Ethnicity, Race & Migration', 'French', 'Italian', 'Judaic Studies', 'Latin American Studies', 'Lit. and Comparative Cultures', 'Modern Middle Eastern Studies', 'Near Eastern Languages & Civs', 'Religious Studies', 'Russian & E European Studies', 'Spanish', 'Special Divisional Major',
"Women'sGender&SexualityStudies"

Category 4: Life and Environmental Sciences

'Cognitive Science', 'Ecology & Evolutionary Biology', 'Environmental Engineering', 'Environmental Studies', 'Geology & Natural Resources', 'Molecular Biophysics & Biochem', 'Molecular,Cellular,Dev Biology', 'Molecular,Cellular,DevBio(Int)', 'Neuroscience', 'Physics & Geosciences'

Category 5: Social Sciences, Economics, Political Science

'Anthropology', 'Archaeological Studies', 'Economics', 'Economics & Mathematics', 'Global Affairs', 'Global Health', 'Linguistics', 'Political Science', 'Political Science (Int.)', 'Psychology', 'Sociology', 'Sociology (Int.)'

Category 6: Undeclared

‘Undeclared'