

# DISCUSSION PAPER SERIES

DP17809

## **THE DYNAMICS OF NETWORKS AND HOMOPHILY**

Matthew O. Jackson, Stephen Nei, Erik Snowberg  
and Leeat Yariv

**DEVELOPMENT ECONOMICS,  
INDUSTRIAL ORGANIZATION, LABOUR  
ECONOMICS AND POLITICAL  
ECONOMY**

**CEPR**

# THE DYNAMICS OF NETWORKS AND HOMOPHILY

*Matthew O. Jackson, Stephen Nei, Erik Snowberg and Leeat Yariv*

Discussion Paper DP17809  
Published 14 January 2023  
Submitted 24 December 2022

Centre for Economic Policy Research  
33 Great Sutton Street, London EC1V 0DX, UK  
Tel: +44 (0)20 7183 8801  
[www.cepr.org](http://www.cepr.org)

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Development Economics
- Industrial Organization
- Labour Economics
- Political Economy

Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

Copyright: Matthew O. Jackson, Stephen Nei, Erik Snowberg and Leeat Yariv

# THE DYNAMICS OF NETWORKS AND HOMOPHILY

## Abstract

We examine friendships and study partnerships among university students over several years. At the aggregate level, connections increase over time, but homophily on gender and ethnicity is relatively constant across time, university residences, and different network layers. At the individual level, homophilous tendencies are persistent across time and network layers. Furthermore, we see assortativity in homophilous tendencies. There is weaker, albeit significant, homophily over malleable characteristics---risk preferences, altruism, study habits, and so on. We find little evidence of assimilation over those characteristics. We also document the nuanced impact of network connections on changes in Grade Point Average.

JEL Classification: D85, I21, J15, J16, Z13

Keywords: Homophily, Social Networks

Matthew O. Jackson - [jacksonm@stanford.edu](mailto:jacksonm@stanford.edu)  
*Stanford University*

Stephen Nei - [s.m.nei@exeter.ac.uk](mailto:s.m.nei@exeter.ac.uk)  
*University Of Exeter*

Erik Snowberg - [snowberg@eccles.utah.edu](mailto:snowberg@eccles.utah.edu)  
*University Of Utah*

Leeat Yariv - [lyariv@princeton.edu](mailto:lyariv@princeton.edu)  
*Princeton University and CEPR*

# The Dynamics of Networks and Homophily\*

**Matthew O. Jackson**  
Stanford University,  
and the Santa Fe Institute  
jacksonm@stanford.edu  
[stanford.edu/~jacksonm](https://stanford.edu/~jacksonm)

**Erik Snowberg**  
University of Utah,  
UBC, CESifo, and NBER  
snowberg@eccles.utah.edu  
[eriksnowberg.com](https://eriksnowberg.com)

**Stephen M. Nei**  
University of  
Exeter  
s.m.nei@exeter.ac.uk  
[sites.google.com/site/sneissite/](https://sites.google.com/site/sneissite/)

**Leeat Yariv**  
Princeton University,  
CEPR, CESifo, and NBER  
lyariv@princeton.edu  
[lyariv.mycpanel.princeton.edu](https://lyariv.mycpanel.princeton.edu)

December 23, 2022

## Abstract

We examine friendships and study partnerships among university students over several years. At the aggregate level, connections increase over time, but homophily on gender and ethnicity is relatively constant across time, university residences, and different network layers. At the individual level, homophilous tendencies are persistent across time and network layers. Furthermore, we see assortativity in homophilous tendencies. There is weaker, albeit significant, homophily over malleable characteristics—risk preferences, altruism, study habits, and so on. We find little evidence of assimilation over those characteristics. We also document the nuanced impact of network connections on changes in Grade Point Average.

JEL CLASSIFICATIONS: D85, I21, J15, J16, Z13

KEYWORDS: Homophily, Social Networks, Dynamic Networks, Undergraduate Education, Peer Effects

---

\*We gratefully acknowledge financial support under ARO MURI Award No. W911NF-12-1-0509 and NSF grants SES-1629446 and SES-2018554.

# 1 Introduction

Network structure influences a range of behaviors and economic outcomes: product choices, political decisions, and labor-market opportunities, and many others.<sup>1</sup> *Homophily*—the tendency to connect with individuals who are similar in some way—is often observed in network structures.<sup>2</sup> Homophily mediates a variety of network effects, including the speed at which behavior and information spreads, the opportunities individuals have access to, as well as the formation of behavioral norms and culture.<sup>3</sup> Despite homophily’s importance, the paucity of longitudinal data on this phenomenon leaves many important questions unanswered: Is homophily persistent over time and does it vary across network layers? Is it an individual trait or an aggregate network property? Is there assortativity in homophilous tendencies? Is homophily on malleable characteristics the product of assimilation or selection? Is homophily over certain characteristics and network layers particularly important for outcomes?

We answer these questions in the context of an undergraduate student population, where we trace the evolution of friendships and study partnerships at the individual level over three years. There are several main findings. First, at the network level, homophily on both gender and ethnicity appear early and are stable over time and across layers—friendship and study partnerships, in particular. Second, homophilous tendencies on gender and ethnicity differ across individuals and, at the individual level, are consistent across network layers and persistent over time, suggesting that homophily is a stable trait. Third, we see homophily over homophily: more homophilous individuals tend to connect with more homophilous others. Fourth, there is homophily on malleable characteristics (for example, altruism, risk aversion, hours playing video games), but it appears with a lag. Moreover, there is little evidence of assimilation on malleable characteristics: homophily appears to be driven predominantly by the selection of connections. Fifth and finally, homophily has important, but nuanced,

---

<sup>1</sup>See [Jackson \(2010\)](#), [Bramoullé et al. \(2016\)](#), and [Jackson et al. \(2017\)](#) for discussion and references.

<sup>2</sup>Homophily may be the result of exposure, preferences, differences in the average numbers of friendships that different groups have, and so on. In our use of the term, homophilous connections do not necessarily imply homophilous preferences.

<sup>3</sup>See [Verbrugge \(1977\)](#), [McPherson et al. \(2001\)](#), [Currarini et al. \(2009\)](#), and [Goeree et al. \(2010\)](#) for background and references on homophily. For examples of its impact on behaviors, see [Zhuravskaya et al. \(2020\)](#), [Jackson \(2021\)](#), and [Chetty et al. \(2022a\)](#).

impacts on outcomes. It operates differently through study partnerships and friendships. In particular, ethnic homophily in study partnerships slightly reduces students' Grade Point Average (GPA) gains, while gender homophily leads to GPA gains, but only for females.<sup>4</sup>

Our study is enabled by a unique longitudinal data set—the Caltech Cohort Study (CCS), described in Section 2. The CCS combines four extensive, incentivized surveys over three years, starting from the fall of 2013. Each survey was taken by roughly 90% of the student body at the California Institute of Technology (Caltech). Thus, we have rich data on individual's choices across time. The surveys elicited students' friendships and study partnerships, in addition to a battery of malleable characteristics: risk aversion, altruism, over-confidence, over-precision, implicit attitudes toward gender and race, and so on. We wed the survey data with institutional data on students' demographic characteristics, housing arrangements, and academic performance. We focus on the population of students entering Caltech in 2013 and follow them over three years.

Homophily in gender and ethnicity is pronounced and relatively stable across both friendships and study partnerships, as shown in Section 3. Homophilous links appear at a significantly higher frequency, and the more attributes individuals have in common, the higher the chance that they are linked. The aggregate frequency of homophilous and non-homophilous links both change over time, but in relative synchronicity. Although gender and ethnic homophily are substantial in both the friendship and study partnership networks, gender homophily is more pronounced in the friendship network.

Students also exhibit considerable homophily with those who live in the same house or share an academic major, although this does little to alter patterns of gender and ethnic homophily. At the time of our study, students' housing assignments responded to their preferences, as well as the preferences of previous student cohorts already in the houses, through a matching protocol that took place during the first few weeks after students arrived. Thus, it is unsurprising that students have a strong tendency to form connections with

---

<sup>4</sup>In the U.S., students are graded on an A–F scale (omitting E). Typically, an A is worth 4 points, a B is worth 3 points, ..., and an F is worth 0 points. A GPA averages over these scores, weighted by the number of credits received for a class, and is an indicator of overall academic performance.

people in their own house. To a limited extent, this homophily in housing crowds out some ethnic homophily: ethnic homophily is slightly lower for in-house rather than out-of-house connections. Overall, however, housing tends to concentrate homophilous connections rather than diversify them.

Individual-level homophilous tendencies are consistent across network layers and stable over time, as shown in Section 4. There is substantial heterogeneity in individual-level homophilous tendencies underlying the network-level homophily patterns we observe. However, individual tendencies are consistent across network layers and across time: individuals with a larger fraction of friends of a similar gender or race tend to have more similar study partners as well, and the profile of their friends and study partners is stable over time.

Students exhibit considerable assortativity on homophily itself, as shown in Section 5. While there is limited variation in the distribution of genders and ethnicities across houses, homophilous tendencies vary substantially by house. This aggregate pattern largely emerges from students' tendency to form connections with others that have a similar level of homophilous connections, both at the house level—some houses have more homophilous individuals than others—and, at the individual level, homophilous tendencies are correlated across friends. We assess the significance of these patterns using a new simulation method based on a technique introduced by Fosdick et al. (2018). The method allows us to simulate random networks respecting arbitrary constraints: the number of links each student has within and outside their house, the number of same-gender or same-ethnicity links each student has, and so on. The resulting simulations allow us to account for mechanical forces that might generate the observed assortativity in homophily, and show that the assortativity is in excess of what is generated structurally. That is, we see homophily over homophily in our data.

We document homophily on many malleable characteristics elicited through the survey. Similarity among linked individuals on such characteristics could, in principle, be an artifact of either selection of similar connections or a process of assimilation occurring once connections are formed. Homophily over malleable characteristics appears to be largely the result

of selection rather than assimilation, as we show in Section 6. In particular, while there is some reduction in distance between the behavioral attitudes of individuals and their friends and study partners over time, it is extremely rare for connections retained across successive years to exhibit greater similarity on any of these attitudes. That is, replacement of previous friends or study partners with ones who are more similar explains any increased resemblance on malleable characteristics between linked individuals.

One exception, where there appears to be some assimilation in specific circumstances, comes from our analysis of changes in GPA, in Section 6.3. Student GPAs increase when they have a study partnership with a stronger student, but only if both members of the study partnership are women. These patterns are not present for male-only study partnerships, nor mixed-gender partnerships, nor for friendships that are not also study partnerships. In particular, while homophily patterns are generally similar across the friendship and study partnership networks, the two layers have different implications when it comes to outcomes. This highlights the importance of considering various layers of interactions when assessing network effects.

Our results have important implications for labor market skills gained through college (see, for instance, Brewer and McEwan, 2010; Heckman and Mosso, 2014). Our results suggest that by the time individuals reach college, homophily is ingrained: different exposures, experiences, or even time do not alter homophilous tendencies in meaningful ways. In particular, engineering social and academic connections—especially through limiting easy access to certain others, as on-campus residential arrangements do—can be a challenging task. In our study, while students tend to connect with others in their house and in their major, those tendencies do not significantly alter the homophilous features of their friends and study partners.

We hope the array of stylized facts we document provides guidance for further development of dynamic network formation models. Existing models largely focus on a single dimension on which individuals may be similar. Our study demonstrates that relationships between multiple dimensions of similarity, type-dependent severance of links, assortativity



on homophily, and learning of others’ features over time, are all important in explaining homophily.<sup>5</sup> Refining current models can inform the design of policies that take anticipated peer effects into account.

**Related Literature** Homophily has been demonstrated over a rich set of attributes—political affiliations, demographic backgrounds, appearance, and so on—across a diverse set of literatures in economics, political science, sociology, and psychology (see, for example [Verbrugge, 1977](#); [McPherson et al., 2001, 2021](#)).

Detailed longitudinal data sets allowing for the assessment of homophily within full communities are rare. Nonetheless, several empirical studies have inspected the patterns of homophily on sampled data. For example, [Shrum et al. \(1988\)](#) study a sample of children between grades 3 and 12 and document gender and racial homophily. [Pearson et al. \(2006\)](#) use the Teenage Friends and Lifestyle Study to track 160 West Scotland students between the ages of 13 and 15, and report homophily and assimilation patterns for drug abuse. [Overgoor et al. \(2020\)](#) use Facebook data to investigate how homophily—in seniority, gender, and place of origin—is affected by different college features such as size, presence of Greek life, and so on. Our study contributes to this literature by analyzing a data set that is unique in two respects. First, we observe networks within a nearly full population of students, over several layers—friendships, study partnerships, and housing—and over multiple years. Second, we observe a broad set of individual attributes through both institutional data and repeat incentivized surveys. The unique features of our data allow us to paint a rich picture of the patterns and dynamics of homophily, over a wide range of attributes and across network layers, examining their interactions and their implications for scholastic outcomes.

While many studies elicit friendship connections (see, for instance, [Jackson, 2010](#)), relatively few explicitly consider study partnerships or geographical proximity, and rarely in

---

<sup>5</sup>For a review of early network formation models, see [Jackson \(2005\)](#). For more recent models of homophily see [Currarini et al. \(2009\)](#), [Baccara and Yariv \(2013\)](#), [Song and van der Schaar \(2015\)](#), [Graham \(2016\)](#), [Zuckerman \(2022\)](#), and references therein. These models hinge on agents’ evolving valuations of links, which our results speak to directly. Of particular note, [Fu et al. \(2012\)](#) study an evolutionary model generating homophily allowing for multiple phenotypes, and [Graham \(2016\)](#) studies identification of homophily and clustering preferences, allowing for homophily to operate over multiple dimensions.

conjunction, as we do. [Sacerdote \(2001\)](#) studies the impacts of random dorm assignments of Dartmouth students on academic performance and fraternity memberships. [Burns et al. \(2015\)](#) and [Garlick \(2018\)](#) use a sample of students from the University of Cape Town and show that roommates’ racial and academic profiles affect prejudice, volume of interracial interactions, and GPAs.<sup>6</sup> The multi-layer network we observe enables us to isolate the layers that homophily operates on. In particular, while we see strong homophily on housing, in line with prior work, the ethnic and gender homophily we observe among friends and study partners is orthogonal to geographical proximity. The link between homophilous tendencies and diversity of connections has been suggested by [Somashekhhar \(2014\)](#) in the context of job referrals, albeit via different patterns than those we identify.<sup>7</sup>

Our data allow us to study *individual-level* relationships between homophilous tendencies across different layers of the network, and the persistence of those tendencies over time. Much of the prior literature on homophily focuses on aggregate patterns. Trends in population-level homophily still allow individuals to exhibit varying homophilous tendencies across different network levels and over time. Our detailed data allow us to document a strong relationship between an individual’s homophilous tendencies on the two network layers we focus on: friendships and study partnerships. Moreover, these tendencies are stable over time: individuals with more homophilous links at one point in time have more homophilous links at other points in time, even when accounting for relative trait frequencies.

There is little research examining whether there is assortativity in, or homophily over, homophily itself. The assortativity in homophilous tendencies we report is different from the “homophily paradox” defined by [Evtushenko and Kleinberg \(2021\)](#). They offer a theoretical argument for why any homophilous link is more likely to occur between more homophilous individuals. We show that, in excess of this mechanical feature, homophilous tendencies are strongly assortative.

---

<sup>6</sup>[Baccara et al. \(2012\)](#) consider three different social and academic network layers to explain behaviors of university faculty in the context of geographical office assignments.

<sup>7</sup>In particular, [Somashekhhar \(2014\)](#) indicates that homophilous tendencies among minorities *within* the workplace lead to non-minority colleagues being exposed to minorities *outside* the workplace. We show that, in addition, when there is homophily over multiple dimensions, homophily and diversity of connections can coexist even *within* a small social unit (in our case, houses).

Our effort to disentangle selection from assimilation as a channel for homophily in behavior and malleable characteristics has several antecedents, mostly in the context of substance abuse among students: see, for instance, [Kandel \(1978\)](#), [De La Haye et al. \(2013\)](#), [Veenstra et al. \(2013\)](#), [Osgood et al. \(2013\)](#), and [Barnett et al. \(2022\)](#).<sup>8</sup> These papers commonly consider one network layer, a limited set of attributes, and often a far smaller set of individuals than we consider. Consequently, results emerging from this literature are mixed, with some finding evidence of assimilation, and others of assortativity. In our setting, assortativity consistently dominates across a range of behaviors and over multiple network layers. In recent years, the literature has developed many econometric techniques for disentangling selection from assimilation (see Chapter 8 in [Zafarani et al. 2014](#) and [Snijders 2017](#) for reviews). The longitudinal nature of our data allows us to use simple variants of these techniques.

Our results on the effects of homophily on students' GPA complement a large literature on peer effects in academic achievement. For example, [Sacerdote \(2001\)](#) takes advantage of random assignment of students to dorms at Dartmouth and identifies peer effects relating to GPA at the dorm-room level. [Garlick \(2018\)](#) compares changes in GPA of University of Cape Town students who are assigned to dorms randomly to those grouped based on past academic performance. Grouping low-GPA students leads to worse outcomes, while grouping high-GPA students does not have a significant benefit. In contrast, our results on the effects of study partnerships echo those of [Cools et al. \(2019\)](#), who find that grade-school girls improve performance when grouped with high-achieving girls, defined in the study as those with at least one parent with a graduate education. However, girls' performance worsens when grouped with high-achieving boys. Boys' performance is unaffected by their grouping. These studies do not distinguish between social and scholastic interactions, and use limited demographic and behavioral controls. Our distinct elicitations of friendships

---

<sup>8</sup>[de Klepper et al. \(2010\)](#) document assimilation of military discipline across friendships between Dutch naval officers, and [Bhargava et al. \(2022\)](#) estimate significant peer effects on several behavioral traits in friendship networks within high-school classrooms at one point in time. In the context of product adoption, [Aral et al. \(2009\)](#) use data from a global instant messaging platform to show that selection plays an important role relative to diffusion. Some studies focus only on assortativity based on difficult to observe behavioral traits. For instance, [Brañas-Garza et al. \(2022\)](#) study 168 University of Granada freshmen and report limited friendship assortativity based on behavioral traits.

and study partnerships allow us to identify the type of connections that produce the most pronounced effect on GPA—in our setting, study partnerships. In addition, we account for a broad set of demographic and behavioral attributes when assessing GPA dynamics.<sup>9</sup>

## 2 The Caltech Cohort Study

Caltech is an independent, private university located in Pasadena, California. It has around 900 undergraduate students. Its size and relatively closed community enable us to get a fairly comprehensive look at the relationships among a cohort over a period of several years.

In the fall of 2013, 2014, 2015, and the spring of 2015, we administered an incentivized online survey to the entire undergraduate student body. We elicited students' friendships by asking students to name five of their closest friends in each installment of the survey. In the fall of 2014 and 2015, we also elicited students' five closest study partners. In addition, we used incentivized tasks to elicit an array of attributes, including risk aversion, altruism, ambiguity aversion, competitiveness, cognitive sophistication, honesty, overconfidence, overprecision, optimism, and implicit attitudes toward gender and race. Students were also asked a large set of questions addressing their lifestyle and social habits, including their sleep patterns, study routines, and physical attributes.<sup>10</sup>

Most of our analysis focuses on the fall surveys in order to consider changes over similar intervals of time. In the fall of 2013, 88% of the student body (806/916) responded to the survey. The average payment was \$20.58. In the fall of 2014, 92% of the student body (893/972) responded to the survey, and the average payment was \$24.34. Of those who took the survey in 2013 and did not graduate, 89% (546/615) also took the survey in the fall of 2014. In the Fall of 2015, 87% (875/1,005) of the student body responded to the survey. The

---

<sup>9</sup>Carrell et al. (2013) connect the endogenous formation of study groups with outcomes. They use imposed squadrons in the US Air Force Academy to study peer effects on lower-achieving students' performance. The endogenous emergence of study sub-groups leads lower-achieving students to exhibit *lower* performance in the imposed squadrons. Corno et al. (2022) document the beneficial effects of interracial dorm allocations on Black South African college students.

<sup>10</sup>Sample survey screenshots are available at: [lyariv.mycpanel.princeton.edu/~papers/ScreenshotsFall2014.pdf](https://lyariv.mycpanel.princeton.edu/~papers/ScreenshotsFall2014.pdf). The AEA RCT registry was launched in 2012. At the time our surveys were run, pre-registration was not common for non-RCT studies.

Table 1: Summary statistics

|                         | 2013 Cohort | 2010–2015 Cohorts |
|-------------------------|-------------|-------------------|
| Panel A: Gender         |             |                   |
| Female                  | 34.77%      | 38.43%            |
| Male                    | 65.23%      | 61.57%            |
| Panel B: Ethnicity      |             |                   |
| Asian                   | 46.88%      | 43.13%            |
| Black                   | 1.56%       | 1.34%             |
| Caucasian               | 27.34%      | 29.11%            |
| Hispanic                | 9.77%       | 10.93%            |
| International           | 10.16%      | 9.73%             |
| Two +                   | 4.30%       | 5.50%             |
| Panel C: House Clusters |             |                   |
| South                   | 42.97%      | 43.26%            |
| North                   | 39.06%      | 39.50%            |
| Far North               | 17.19%      | 16.03%            |
| Other                   | 0.78%       | 1.21%             |
| Panel D: GPA            |             |                   |
| GPA (2015)              | 3.47        | —                 |

average payment was \$29.25. Of those who took the survey in 2014 and did not graduate, 87% (621/712) also took the survey in the fall of 2015. Unlike most surveys, there is little concern about self-selection into ours due to the high response rates; see [Snowberg and Yariv \(2021\)](#) for an analysis of selection into the CCS.

In addition to the survey data, we have institutional data on all students’ academic outcomes and demographics—including gender, race, country of origin, major, and grade point average (GPA) throughout their time at Caltech. We also have data on students’ college housing. Caltech’s undergraduates have the option of living in one of the eight residential houses on campus, which are divided into three geographical clusters: North, South, and Far North. Houses within a cluster exhibit strong connections in terms of proximity and shared

activities. Although the allocation to on-campus houses is not random, we use housing data in some of our analyses as controls. Our analysis focuses on the cohort entering in the fall of 2013; we track those students from their arrival at Caltech. Table 1 provides summary statistics for this cohort, as well as for the students that members of the 2013 cohort might connect with—those in the cohorts of 2010–2015, whose years at Caltech overlap with the 2013 cohort.

Caltech is a highly selective institution, which may cause one to worry that its students are not representative of students more broadly. Several considerations should alleviate such concerns. First, in many ways, Caltech students behave similarly to other student populations. Responses from our survey to several standard elicitations—of risk, altruism in the dictator game, and so on—are similar to those reported in several other pools (see [Gillen et al. 2019](#) and [Snowberg and Yariv 2021](#)).<sup>11</sup> Second, although different on-campus student populations have their idiosyncracies, they all feature the undergraduate experience as a unique period in which important life decisions are made, new friendships are formed, and both social and scholastic interactions occur within a contained environment. These common features make student populations particularly interesting for the focus of this study: how interaction networks evolve. Last, the student bodies of highly selective institutions are substantial, and of interest in their own right: in the US alone, the college-age population attending top-50 schools accounts for about a million students, many of whom go on to leadership positions.<sup>12</sup>

### 3 Dynamics of Connections

As students acclimate, both the number and profile of their friendships and study partnerships evolve. Despite this churn, students persistently form a disproportionate number of their relationships with others of the same ethnicity, gender, or both. Shared housing as

---

<sup>11</sup>In addition, [Chetty et al. \(2022b\)](#) examine economic homophily across many universities, and find that Caltech exhibits reasonably representative homophily patterns.

<sup>12</sup>This figure is derived from statistics produced by the National Center for Education Statistics and the US News and World Report college rankings data.

well as academic majors increase the likelihood of links, but do not substantially alter the patterns of ethnic and gender homophily.

### 3.1 The Number of Links over Time

Friendships were elicited in all three fall surveys, while we began eliciting study partnerships only in the fall of 2014 after a year-long engagement in classwork. Figure 1 shows the basic dynamics of the number of friends and study partners, breaking out homophilous relationships on ethnicity and gender, and those relationships that are new versus those retained from the previous year.

Specifically, Figure 1 depicts the average number of friendships and study partnerships over time for students entering Caltech in 2013. These reflect directed links: for each student, we calculate the number of others named as a friend or study partner.<sup>13</sup> Reciprocated links are less frequent, but exhibit similar patterns: for links within the 2013 cohort, 53% of both friendships and study partner links are reciprocated across all fall surveys.<sup>14</sup>

As can be seen from the gray, right-most bars of Figure 1, the average number of friendships increases between students' freshman and sophomore years, but remains fairly stable between students' sophomore and junior years. The number of study partnerships declines slightly between students' sophomore and junior years. Most friendships occur within the 2013 cohort.<sup>15</sup> Students have an average of 0.6 fewer study partners than friends, which is also shown in Figure 1. There is still a substantial overlap between friends and study partners. For example, in the fall 2015 survey, 51% of the 2013-cohort students' study partners are also friends, and 40% of the 2013-cohort students' friends are also study partners.

Connections exhibit more persistence in later years of students' tenure. The percent of friendships formed during students' freshman year that are still present in their sophomore

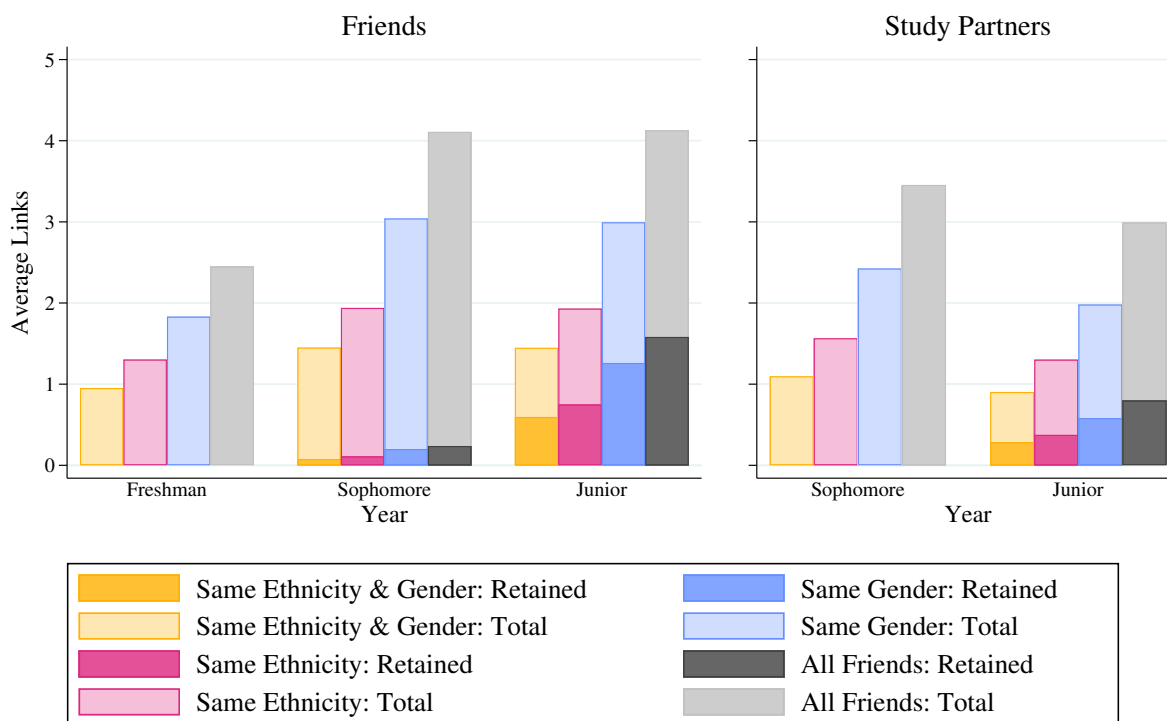
---

<sup>13</sup>In standard terminology, these are out-degrees.

<sup>14</sup>These numbers are similar to the reciprocation rates among 7-12th grade students in Add Health, who were each asked to nominate 10 friends; see, for example, [Vaquera and Kao \(2008\)](#). For reciprocity in other types of networks see, for instance, [Ready and Power \(2021\)](#).

<sup>15</sup>With time, 2013-cohort students form more connections with students in other cohorts. In fall 2013, only 5% of friendships are with students in other cohorts. This percentage increases to 22% in fall 2014 and 28% in fall 2015.

Figure 1: Number of friendships and study partnerships across time, broken out by retained and new relationships.



year is 12%. This figure rises to 36% between sophomore and junior years.<sup>16</sup> Although we observe the 2013-cohort students' study partners only in their sophomore and junior years, just 21.4% of study partnerships observed during students' sophomore year persist in students' junior year, which is a bit less than two-thirds of friendships' survival rate during that period.<sup>17</sup> Finally, friendships that are also study partnerships have the highest likelihood of survival: 42% of such links persist between students' sophomore and junior years.

A substantial fraction of students' links are with individuals of the same ethnicity, gender, or both, as shown by the orange, pink, and blue bars within each year in Figure 1. Despite the high turnover in connections, the number of same-gender and same-ethnicity links increases

<sup>16</sup>These percentages are somewhat higher when considering reciprocated links. The percent of reciprocated friendships formed during a student's freshman year that are still present in their sophomore year is 14%, while the percent of reciprocated friendships observed during a student's sophomore year that are still present in their junior year is 42%.

<sup>17</sup>Reciprocated study partnerships have a higher rate of survival: 28%.



in proportion to the overall number of connections in each survey. Relationships between similar individuals are more likely to be reciprocated: 56% of friendships between students of the same ethnicity or gender are reciprocated, and 58% of friendships between individuals who are the same on both, versus 50% of friendships between students of different ethnicity and gender. Moreover, retention rates are higher for connections between similar individuals. Between freshman and sophomore years, 8–11% of friendships between individuals similar in either gender or ethnicity are maintained, versus 1.6% for those that feature different genders and ethnicities; between sophomore and junior years, the analogous figures are 39–42% versus 29%. There is also a higher percentage of overlapping links among similar students: 43% of same-ethnicity or same-gender friendships are also study partnerships, versus 40% for friendships between individuals that differ in gender or ethnicity. The differences are more pronounced when examining study partnerships: 58–60% of similar study partners are also friends, versus 42% of study partners who are of different genders and ethnicities.

The fact that most relationships are between individuals with similar characteristics is not necessarily evidence of homophily: the distribution of characteristics in the population could, in principle, generate this mechanically. As an extreme example, if all Caltech students were female, 100% of links would occur between same-gender pairs. While the observed student body composition is not as extreme, the question remains whether the similarity patterns we document are mechanical in nature.

### **3.2 Homophily over Gender and Ethnicity**

In this subsection, we examine the scope of homophilous tendencies—in terms of either gender or ethnicity—as well as how they evolve over a student’s tenure. We show that homophily is pronounced across friendships and study partnerships. Homophily levels are similar across these two network layers, although gender homophily is somewhat stronger among friends. Gender homophily is stable over time, while ethnic homophily decreases slightly over students’ first year at Caltech.

To get an initial sense of the extent of homophily—the number of friendships between

similar individuals in excess of what would be expected by chance—we compare the percentage of same-gender and same-ethnicity links in our data with those we would expect to see if students were choosing relationships at random. If all students had exactly the same number of connections, then calculating the expected fraction of homophilous connections given the distribution of attributes in our population would be straightforward.<sup>18</sup> If links were formed at random, and students all had the same number of connections, we would expect 55% of connections to be between individuals of the same gender, contrasting the 76% of same-gender friendships and 69% same-gender study partnerships in our data for the 2013 cohort. Similarly, with no homophily, 32% of links would be between individuals of the same ethnicity, contrasting the 49% of same-ethnicity friendships and 45% same-ethnicity study partnerships in the data.

Some excess same-type links observed in our data may still be generated mechanically by different genders and ethnicities having different numbers of relationships. For instance, in fall 2014, Asian students had an average number of 4.1 friends and 3.3 study partners; Black students had a higher average of 5.0 friends and 4.7 study partners. Similarly, in fall 2014, female students had an average of 3.9 friends and 3.3 study partners, while male students had an average of 4.2 friends and 3.5 study partners. In fact, in the data, gender and ethnicity are associated with a different distribution of numbers of friends and study partners. These differences affect the resulting number of links between similar students even were connections made at random.<sup>19</sup>

We use simulations based on the configuration model (Bender and Canfield, 1978) to estimate the baseline level of homophily that would be expected in our data from random matching. The literature has often relied on Coleman’s (1958) homophily index for this purpose. Coleman’s index normalizes the homophily of a specific group or individual accounting for the underlying distribution of the characteristic in question, in our case gender

---

<sup>18</sup>Suppose a fraction  $f_i$  of students are of “type”  $i$ —capturing gender, ethnicity, their interaction, and so on—and there are  $n$  types in the population. With identical number of friends across types, we would expect a fraction  $\sum_{i=1}^n f_i^2$  of directed links to be homophilous.

<sup>19</sup>For example, if each type  $i$  student names  $k_i$  friends, and a fraction  $f_i$  of the population  $k_i$  is of type  $i$ , we would expect a fraction  $\sum_{i=1}^n (f_i k_i) f_i / \left( \sum_{j=1}^n f_j k_j \right) = \frac{\sum_{i=1}^n f_i^2 k_i}{\sum_{i=1}^n f_i k_i}$  of links to point to similar individuals.

or ethnicity. However, Coleman’s index approach does not account for differing numbers of connections across individuals who simultaneously belong to multiple different groups, such as different genders, ethnicities, college residences, and academic majors. Thus, we instead use a version of the configuration model to randomly simulate links without conditioning on characteristics, but respecting the number of friends each individual has.<sup>20</sup>

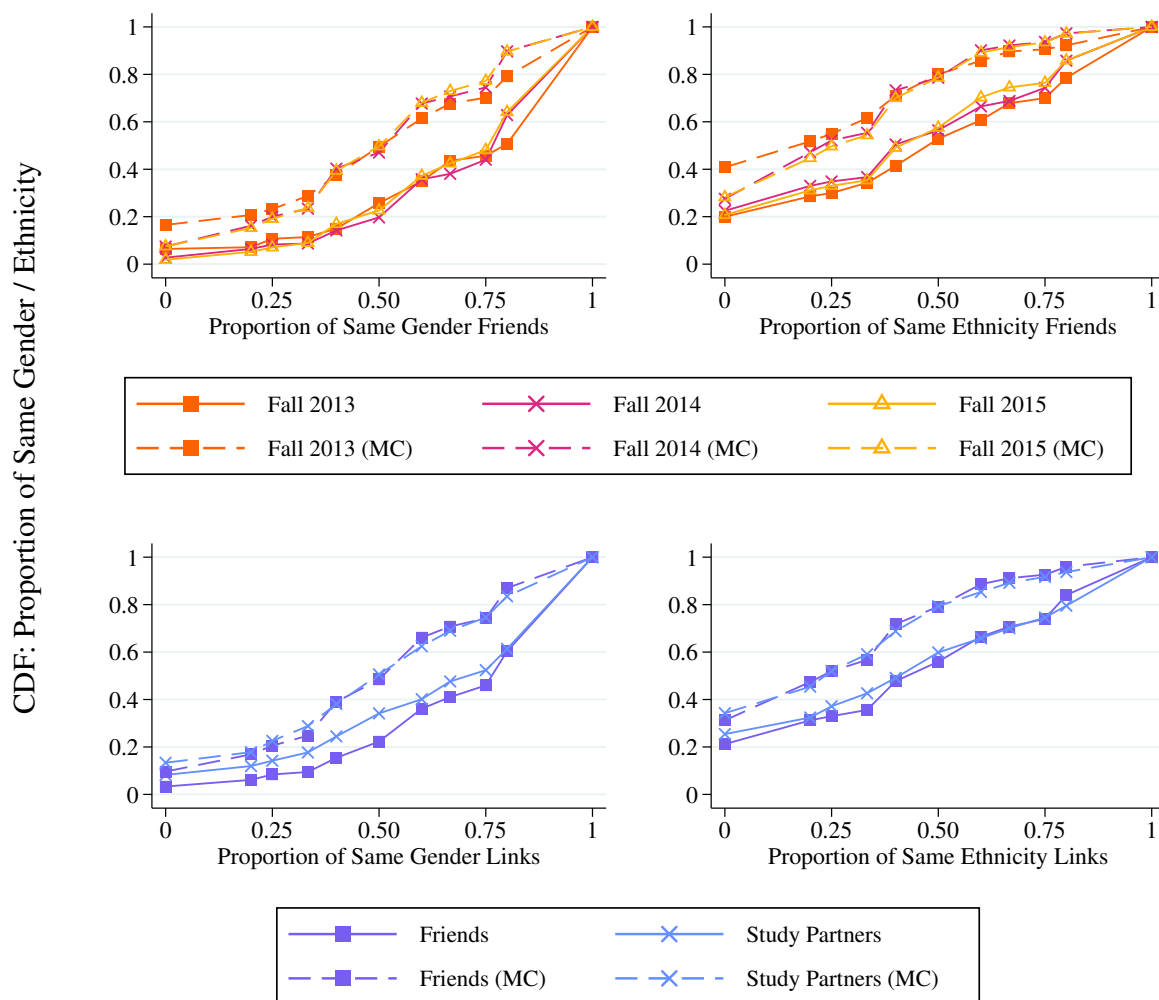
When we account for both the underlying distribution of characteristics and the number of links across different genders and ethnicities, we still find strong evidence for homophilous connections. Our simulations predict essentially the same numbers as those derived assuming a uniform number of links across students: 55% of friendships or study partnerships between those of the same gender, and 32% between those of the same ethnicity. The difference between these simulation results and what we observe in the data occurs across the entire distribution: as Figure 2 illustrates, the distributions corresponding to percentages of same gender and ethnicity friendships first order stochastically dominate the simulated distributions. The figure also illustrates the substantial heterogeneity in the proportion of homophilous connections across individuals, an observation we return to in later sections.

The overall fraction of relationships between individuals of the same gender and ethnicity are relatively stable over time, as illustrated in the top panel of Figure 2. Both the simulated and the observed distributions nearly coincide across all three years. This is despite the fact that there is both an expansion in the number of friends and study partners and high, type-dependent, turnover across waves, as shown in Figure 1. At the aggregate level, gender-based homophily is very stable: 75% of friendships are between individuals of the same gender in the fall of 2013, 74% in fall 2014, and 73% in fall 2015. Ethnic homophily decreases slightly, but the decline is statistically significant ( $p < 0.05$ ) only between fall of 2013 and 2014—decreasing from 53% of friendships being between individuals of the same ethnicity to 47%. The proportion of same-ethnicity friends stays at 47% in fall of 2015. The analogous figures for study partnerships are similar. There is a marginally significant reduction in

---

<sup>20</sup>In the configuration model, one can think of each student as having as many stubs as friends. The simulations connect those stubs at random. This technique allows us to model the random formation of networks, subject to some additional constraints on links.

Figure 2: CDFs of the proportion of friendship links to others of the same gender or ethnicity, broken down by wave, and then layer, with Monte Carlo simulations.



gender homophily between sophomore and junior years, and no significant change in ethnic homophily.<sup>21</sup>

The bottom panel of Figure 2 compares homophily patterns in friendships and study partnerships, using data from all fall surveys. Despite the different numbers of friends and study partners students have, the simulated distributions look virtually identical across the

<sup>21</sup>In fall 2014 and 2015, the fraction of students' same-gender study partners are 70% and 66%, respectively, with the decrease being marginally significant (at  $p < 0.11$ ). The analogous figures for same-ethnicity study partners are 45% and 44%, respectively, which are not significantly different.

two network layers. Ethnic homophily is also quite similar across the two layers in our data. However, the distribution of the proportion of same-gender friends first order stochastically dominates that for study partners, indicating greater gender homophily among friends.

There are channels other than those in the simulations of Figure 2 that may mechanically generate similarities across links. If certain residences or academic majors have a greater percentage of one gender or ethnicity, then homophily in geography or field of study may appear as homophily over gender or ethnicity. Such channels may also account for the stability of observed homophily—students’ houses and majors rarely change over their tenure at Caltech. In the next subsection, we show that although housing and majors indeed affect link formation, controlling for them does not affect the level or statistical significance of homophily in gender and ethnicity. These patterns suggest that, while housing and academic majors may concentrate connections, they may not change the gender and ethnic diversity of connections.

### 3.3 Accounting for Geographical and Academic Proximity

Gender and ethnic homophily are statistically significant and substantial in magnitude, even when controlling for similarity in residence and major, as shown in Table 2. Further, the effects of ethnicity and gender on link formation within each network layer are comparable.<sup>22</sup> The table displays fixed-effect regression results explaining the presence of a directed link between two students through the similarity between them, including dummies for whether the two students are of the same gender, the same ethnicity, or both. For readability, we scale the frequency of links by 1,000.<sup>23</sup>

Houses are a prominent feature of undergraduate student life at Caltech, and the literature (for example, Nahemow and Lawton, 1975; Sacerdote, 2001; Small and Adler, 2019) indicates that proximity is consequential for network formation. At the time of our study,

---

<sup>22</sup>As gender and ethnic homophily do not change much over time, we examine an average over all years here. Table A.2 in the Appendix presents a detailed breakdown of link frequencies by years.

<sup>23</sup>There are close to 1,000 Caltech undergraduates at any point in time. Thus, this normalization implies that coefficients can be roughly interpreted in terms of numbers of links, or “degrees.”

Table 2: Housing and Homophily (per 1,000 potential links)

|                                      | Friends           |                    | Study Partners     |                   |                    |                    |                    |
|--------------------------------------|-------------------|--------------------|--------------------|-------------------|--------------------|--------------------|--------------------|
| Same Gender                          | 3.00***<br>(0.15) | 2.99***<br>(0.15)  | 1.02***<br>(0.19)  | 1.39***<br>(0.12) | 1.35***<br>(0.12)  | 0.36**<br>(0.15)   |                    |
| Same Ethnicity                       |                   | 3.26***<br>(0.17)  | 1.16***<br>(0.26)  |                   | 1.59***<br>(0.13)  | 0.50**<br>(0.21)   |                    |
| Same House                           |                   | 15.84***<br>(0.22) | 8.13***<br>(0.40)  |                   | 10.84***<br>(0.18) | 6.68***<br>(0.32)  |                    |
| Same Major                           |                   | 2.54***<br>(0.18)  | 2.53***<br>(0.18)  |                   | 4.47***<br>(0.14)  | 4.47***<br>(0.14)  |                    |
| Same Gender and Ethnicity            |                   |                    | 1.01***<br>(0.35)  |                   |                    | 0.32<br>(0.28)     |                    |
| Same Gender and House                |                   |                    | 8.81***<br>(0.55)  |                   |                    | 4.86***<br>(0.43)  |                    |
| Same Ethnicity and House             |                   |                    | 2.49***<br>(0.69)  |                   |                    | 1.36**<br>(0.55)   |                    |
| Same Gender, Ethnicity,<br>and House |                   |                    | 12.45***<br>(0.94) |                   |                    | 6.36***<br>(0.75)  |                    |
| Constant<br>(Non-Homophilous)        | 1.98***<br>(0.11) | 2.60***<br>(0.09)  | -1.45***<br>(0.13) | 1.51***<br>(0.09) | 1.77***<br>(0.07)  | -1.22***<br>(0.10) | -0.50***<br>(0.12) |

Notes: \*\*\*, \*\*, \* denote statistical significance at the 1%, 5%, and 10% level with standard errors in parentheses.

Caltech’s housing system consisted of eight independent houses.<sup>24</sup> Nearly all students in our data live in one of these houses their first year at Caltech, and almost all stay in these houses for a large fraction, or all, of their time at Caltech; only 8.5% of sophomores and 19% of juniors live in Caltech-owned off-campus housing or in housing unrelated to Caltech. Many of these retain an affiliation with one of the campus houses and continue dining and socializing within the housing system.

Similarity in housing and academic major are important predictors of link formation, as shown in Columns 3 and 7 of Table 2, in line with the association between local geography and friendships in prior literature. In fact, the housing variables usually have the largest coefficients across the specifications in Table 2—significantly larger than the coefficients corresponding to a shared major.

Despite their importance in link formation, the inclusion of variables capturing similarity in residence and academic major does not diminish the impact of similarity in gender and ethnicity on link formation.<sup>25</sup> The reason is apparent from the underlying data. For example, 74% of friendships within the same house are between individuals of the same gender, and the percentage of same-gender friendships outside an individual’s house is identical—74%. Altogether, Table 2 suggests that housing and field of study concentrate homophily, rather than limit it.

There are significant complementarities between gender, ethnicity, and housing in the frequency of link formation, as shown in Columns 4 and 8 of Table 2. In particular, the coefficients on the interactions between gender, ethnicity, and housing are all large and statistically significant in Table 2. In what follows, we investigate how these complementarities in linkage frequencies relate to the diversity of students’ friendships.

---

<sup>24</sup>See [en.wikipedia.org/wiki/House\\_System\\_at\\_the\\_California\\_Institute\\_of\\_Technology](https://en.wikipedia.org/wiki/House_System_at_the_California_Institute_of_Technology).

<sup>25</sup>Latent or unobserved groupings could feed into the homophilous tendencies we observe, see Mathews and Volfovsky (2021). We control for housing and field of study (majors) as these are arguably the most salient groupings at Caltech.

### 3.4 Multiple Attributes: Complements or Substitutes?

The significant interaction terms in Columns 4 and 8 of Table 2 imply that there is a much higher frequency of link formation between two individuals of, say, the same gender and house, than between two individuals who are just the same gender. Yet, we also see that the proportion of friends who are of the same gender is the same whether or not those friends are in the same house (at 74%). This can occur because there may be difficulties finding people who match on many attributes simultaneously. This difficulty may lead to greater diversity of friendships when individuals have a smaller pool of friends to choose from.

In order to further examine the interactions between attributes of connected individuals, we consider the full set of friends and study partners reported in the fall surveys. In Table 3, we see mostly insignificant correlations between similarities across different traits. The one exception is the marginally significant *negative* correlation between similarity in housing and major among study partners: study partners who share a major are more likely to reside at different houses. Yet, in columns 4 and 8 of Table 2, the coefficients on all interaction terms are significant, positive, and of substantial magnitude.

Why do we see complementarities in link frequencies, but no reflection of these complementarities in link outcomes? These observations could be an artifact of the underlying attribute distribution and students' search technology. If gender, ethnicity, house location, and major are not strongly correlated, and encounters are random, matching on one of the four would be more likely than matching on more than one. The relative scarcity of individuals who are similar on multiple dimensions, coupled with any cost in forming additional links, would generate such results.<sup>26</sup>

Our analysis here suggests an important message: homophily over multiple attributes, constrained to small populations, can lead to more diverse relationships.<sup>27</sup> However, even

---

<sup>26</sup>In our data, we do not see a significant association between the number of friends or study partners each student has and the individual homophily patterns they exhibit. In particular, students with fewer friends or study partners do not differ significantly in the similarity of their connections.

<sup>27</sup>This observation is reminiscent of what is often referred to as *cross-cutting cleavages* in political science, whereby groups on one cleavage overlap on another, see, for example, Powell (1976), the references therein, and the literature that followed.



Table 3: Correlation between friends' and study partners' traits.

|                | Friends           |                   |                  |            |                   | Study Partners    |                    |            |             |                |            |            |
|----------------|-------------------|-------------------|------------------|------------|-------------------|-------------------|--------------------|------------|-------------|----------------|------------|------------|
|                | Same Gender       | Same Ethnicity    | Same House       | Same Major | Same Gender       | Same Ethnicity    | Same House         | Same Major | Same Gender | Same Ethnicity | Same House | Same Major |
| Same Gender    | 1                 |                   |                  |            | 1                 |                   |                    |            | 1           |                |            |            |
| Same Ethnicity | 0.020<br>(0.021)  | 1                 |                  |            | 0.026<br>(0.026)  |                   |                    |            | 1           |                |            |            |
| Same House     | 0.002<br>(0.021)  | -0.031<br>(0.021) | 1                |            | 0.024<br>(0.026)  | -0.020<br>(0.027) |                    |            | 1           |                |            |            |
| Same Major     | -0.008<br>(0.021) | 0.008<br>(0.021)  | 0.024<br>(0.021) | 1          | -0.027<br>(0.026) | 0.037<br>(0.026)  | -0.043*<br>(0.026) |            | 1           |                |            |            |

Notes: Standard errors from 10,000 bootstrap draws in parentheses. \*\*\*, \*\*, \* denote statistical significance at the 1%, 5%, and 10% level.

with the relatively small size of Caltech’s houses—a student will have approximately 20 cohort-mates in their own house—this constraint results in only a small increase in diversity, an observation we return to in our concluding Section 7.

## 4 Patterns of Individual Homophily

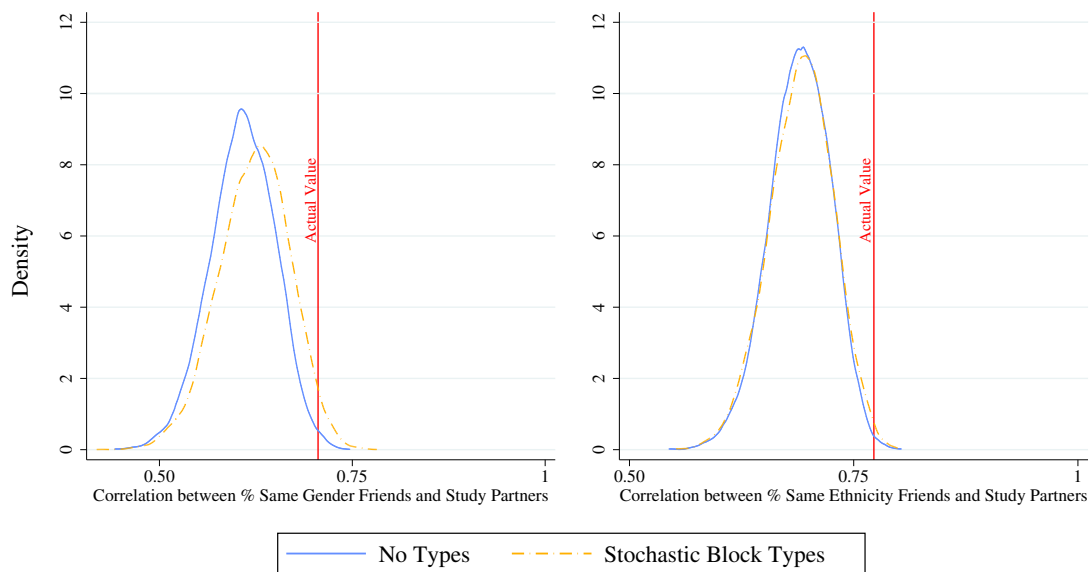
Homophily is clearly an important determinant in the formation of links. However, the analysis in the prior section leaves open the possibility that homophily is a network property—one pertaining to average individual tendencies—rather than a stable attribute of individuals. We now show that the propensity to form homophilous links on gender or ethnicity seems to be largely an individual trait. That is, those with more same gender/ethnicity friends tend to have more same/gender study partners. Furthermore, the fraction of homophilous links early in students’ university experience is persistent over time, despite substantial turnover in the precise identity of friends and study partners.

### 4.1 Individual-level Homophily in Network Layers

The proportion of same gender or same ethnicity links is more highly correlated across friends and study partners than would be expected by chance, as shown in Figure 4. The figure displays the observed correlation between percentages of same gender or ethnicity of friends and study partners, as well as the results of two simulated benchmarks corresponding to random choices of friends and study partners.

First, we consider a benchmark with a *type-free* random formation of links, similar to the simulated model underlying Figure 2. Specifically, given each student’s number of friends and study partners, as well as the number of friends who are also study partners, we simulate these friends’ identities at random from the population 10,000 times. Due to the substantial overlap between friends and study partners, the simulated correlations between the percent of same-gender (or ethnicity) friends and study partners are quite high. However, the observed correlations are higher still, suggesting that those with a greater percentage of friends that

Figure 3: Correlations between percentages of same gender/ethnicity of friends and study partners with no types, stochastic block types.



are of the same gender (or ethnicity) also have a higher percentage of study partners of the same gender (or ethnicity) than would be expected by chance.

Second, we simulate relationships based on a *stochastic block model*. For gender homophily, we estimate the overall fraction  $p_F$  ( $p_M$ ) of female (male) friends and study partners that female (male) students have based on our fall 2014 and 2015 surveys. We then simulate 10,000 networks preserving students' number of friends, where each draw is female with probability  $p_F$  for female students and with probability  $1 - p_M$  for male students. We perform analogous simulations for ethnic homophily, and for the study partnership network layer. Again, the observed correlations between the fraction of same-gender (or ethnicity) friends and the fraction of same gender (or ethnicity) study partners lies well above those generated by friendship and study partnership formation following the stochastic block model.

Figure 3 indicates that while observed correlations between individuals' homophilous tendencies are well below 1—which is expected given the different number of links on each network layer—they exceed those that would be expected by chance. In particular, for gender, the observed correlation is greater than the correlation in 1% of simulations using

the no-types model ( $p < 0.01$ ) and 5% of simulations in the stochastic block model ( $p < 0.05$ ), and, for ethnicity, the observed correlation is greater than the correlation in all but a few simulations ( $p < 0.01$ ). Thus, the similarity in network-level homophily across layers observed in Figure 2 may be driven by consistent individual-level homophilous tendencies.

## 4.2 Persistence of Individual-Level Homophily

To assess whether homophily within a person is stable over time, we focus on friendships.<sup>28</sup> Our analysis in Section 3 illustrated some stabilization of friendship patterns by the beginning of sophomore year. Therefore, we consider friendships that are formed after students have settled into campus and are acquainted with their social environment—that is, those friends named in the fall 2014 and spring and fall 2015 surveys. We code those friends named in fall 2014 as *old* friends, and those added in spring or fall 2015 as *new* friends. The proportion of friends that are of the same ethnicity or gender is more highly correlated across new and old friends than would be expected by chance, and in line with what would be expected if individuals had fixed homophily “types,” as shown in Figure 4.

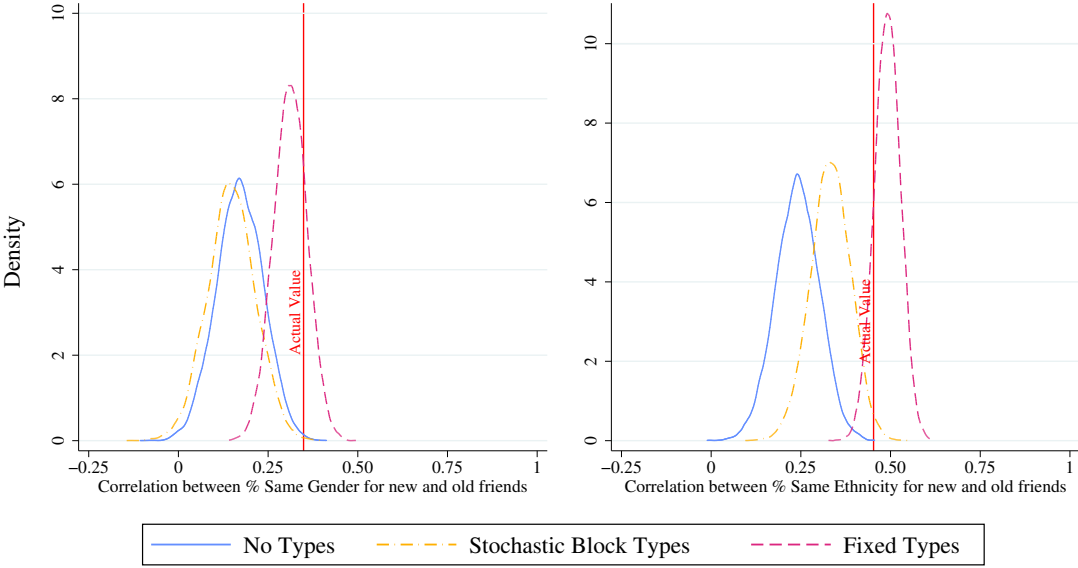
We consider three benchmarks in Figure 4. First, we consider type-free random formation of links as described in the previous subsection. Specifically, given each student’s number of friends, we simulate these friends’ identities at random from the population 10,000 times. The observed correlations between the fraction of same gender (or ethnicity) friends who are old and new lie well above those generated by random friendship formation.

Second, we consider a stochastic block model. As before, for gender homophily, we estimate the overall fraction  $p_F$  ( $p_M$ ) of female (male) friends that female (male) students have based on our fall 2014, spring 2015, and fall 2015 surveys. We then simulate 10,000 networks preserving students’ number of friends, where each draw is female with probability  $p_F$  for female students and with probability  $1 - p_M$  for male students. We calculate the resulting correlation between the fraction of same-gender friends in fall 2014 and at the two observation points in 2015. We perform analogous simulations for ethnic homophily.

---

<sup>28</sup>Results are similar when considering study partnerships, see Appendix Figure A.1

Figure 4: Correlations between percentages of same gender/ethnicity of new and old friends with no types, stochastic block types, and fixed types.



Our third benchmark examines persistence of *homophily types*. Types are represented by the probability that a friend is of the same ethnicity or gender. Under this benchmark, friendship draws are independent. Thus, observed friendships are exchangeable. In our simulations, we preserve the friends we observe for each individual, but randomly draw them, without replacement, into the fall 2014 survey, or into the fall or spring 2015 survey, respecting the number of friends in each survey period. We then compute the correlation between the fraction of same-ethnicity or same-gender friends across those two survey periods. We repeat this random partitioning of friends 10,000 times, and display the resulting correlations between the fraction of old and new friends that are of the same gender or ethnicity.

The type-free random friendship formation model and the stochastic block model generate correlations that are significantly lower than those observed in our data, as shown in Figure 4. Less than 1% of simulations are higher than the observed correlation within each of the four simulated distributions: we can reject these models at the  $p < 0.01$  level.

The fixed-type simulations generate distributions of correlations that have a significant overlap with observed values in the data. Thus, we cannot reject a model in which individuals

have a perfectly persistent type, expressed as the probability that a randomly-drawn friend is of their ethnicity or gender. It is possible to trivially reject more stringent models of types—for example, assuming individuals target, and achieve, the exact same mix of friends over periods—as these would imply correlations equal to 1 over time. The results in Figure 4 suggest persistence in preferences, or the search technology individuals in our sample use in forming friendships.

Taken together, this section provides evidence that homophilous tendencies over gender and ethnicity are individual-level traits. This leads to a natural next question: is there homophily over homophilous tendencies?

## 5 Homophily over Homophily

This section documents substantial sorting over homophily within Caltech’s residential houses. Moreover, we build on a technique for simulating counterfactual network arrangements (based on Fosdick et al., 2018) to develop a new method for assessing assortativity in individual-level homophilous tendencies. We show substantial assortativity, exceeding what would arise from random choices, even when accounting for students’ houses and their friends’ gender and ethnic profile.

### 5.1 Assortativity in Homophily across Houses

During our study period, assignment to Caltech housing was based on the preferences of both the current residents of houses and the incoming students, providing a natural setting for assessing whether more homophilous individuals congregate. Specifically, at the time of the study, assignment to houses was based on a two-sided matching procedure reminiscent of the Gale and Shapley (1962) algorithm.<sup>29</sup> The eight Caltech houses are divided into three geographical clusters: North, South, and Far North houses. Houses within each cluster organize joint events and are in greater proximity to one another. Due to these interactions

---

<sup>29</sup>Thus, unlike assignment of dorms in Dartmouth studied by Sacerdote (2001), initial house assignment at Caltech is not random.

being social in nature, we focus our attention here on friendships. We consider the percentage of same-gender and same-ethnicity friends individuals have throughout our three fall surveys, tracking individuals by demographics and house clusters.

House clusters have different profiles of homophily over gender and ethnicity, as shown in Figure 5. This figure depicts the distributions of percentages of same-gender and same-ethnicity friendships within each housing cluster. It also depicts simulated distributions, broken down by housing cluster. These simulated distributions are generated as in Figure 2: we choose friends at random, but fix the number of friends each student has within their house and other houses to match the level observed in the data. This is done 10,000 times.

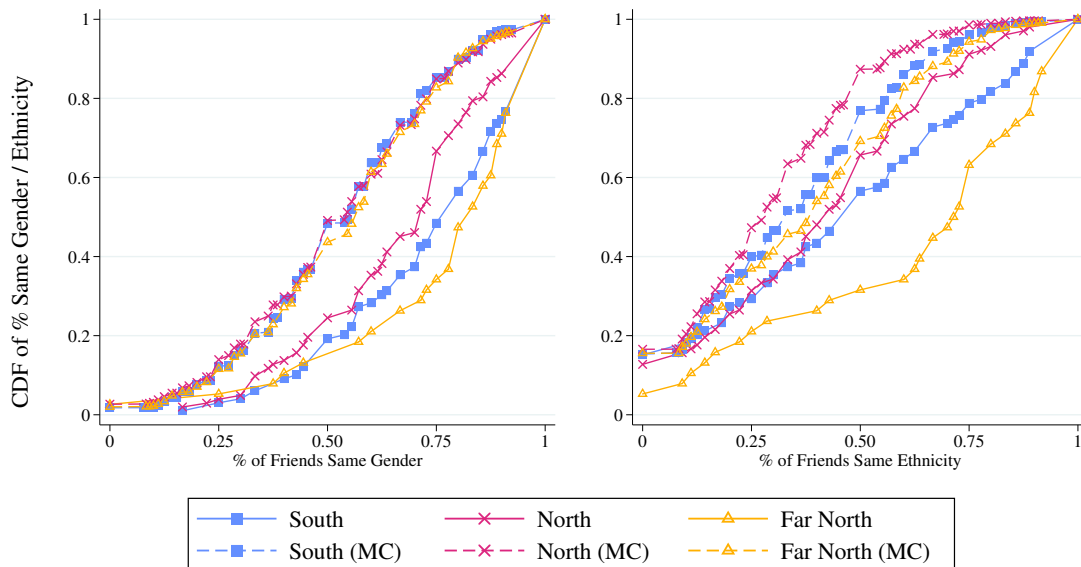
There are substantial differences between the distributions of individual-level homophily across housing clusters, with the Far North house having more homophilous individuals than both the South and North houses. While the overall distributions are not perfectly ordered via first order stochastic dominance, the North houses appear to have more highly-homophilous individuals than the South houses. Given the importance of within-house links identified in Table 2, the observation that homophilous tendencies appear similar within and outside of students' houses, and the fact that student preferences play an important role in housing assignments, this sorting pattern suggests the possibility that homophilous individuals cluster together.

## 5.2 Assortivity in Homophily at the Individual Level

Identifying whether individuals exhibiting high levels of homophily are more likely to have relationships with other high-homophily individuals is challenging: assortativity in homophilous tendencies can be generated by subtle mechanical effects.

[Evtushenko and Kleinberg \(2021\)](#) identify a mechanical force yielding a correlation in homophily within same-type friendships. For example, when considering gender, their results show that female friends of female students are, on average, more homophilous than female friends of male students. Intuitively, individuals with higher homophily are relatively more likely to be the ones connected with similar others. However, this simply implies that the

Figure 5: CDFs of the proportion of friendship links to others of the same gender/same ethnicity, broken down by housing cluster, with Monte Carlo simulations for reference.



Notes: Monte-carlo simulations draw random networks preserving the out-degree of each node. Distributions in the figure result from averaging over 10,000 such random networks.

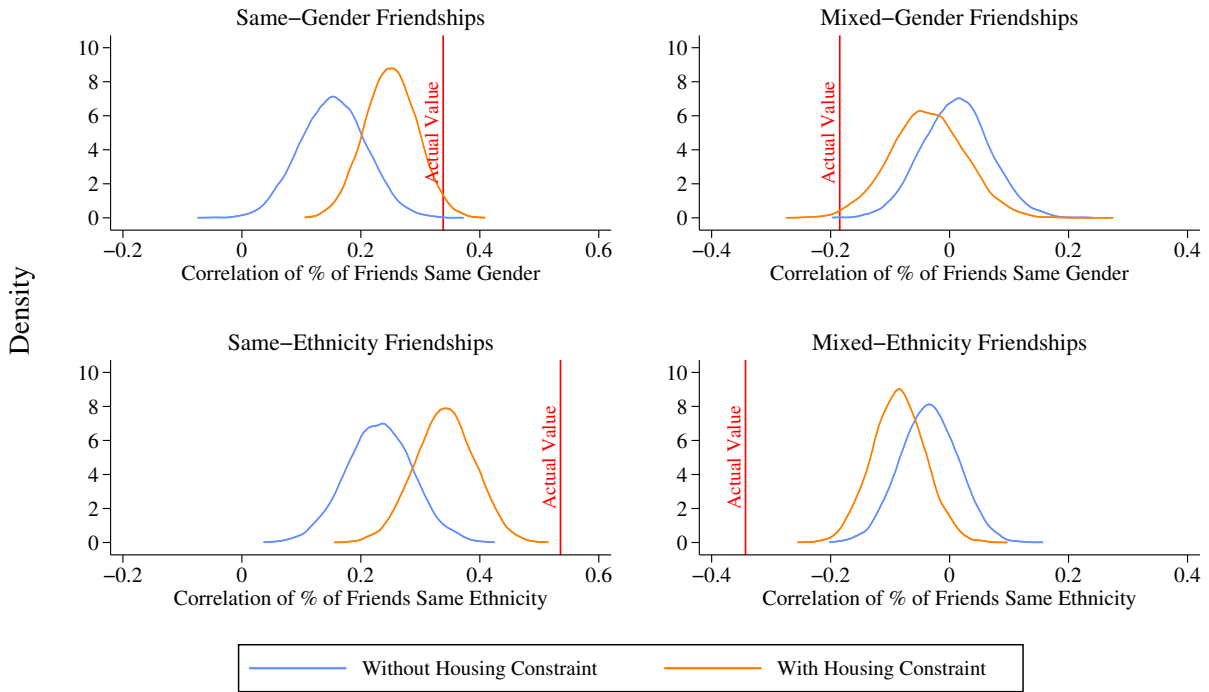
female friends of any female are more likely to be homophilous. This does *not* imply that the female friend of a more homophilous female is more homophilous than the female friend of a less homophilous female. That stronger property is the *assortativity in homophily* that we examine (and find) here.

To analyze the statistical significance of assortativity in homophily, we use simulations. We use a variation of the canonical *configuration model* from random graph theory to generate random networks that satisfy features of the observed network. Namely, focusing on the 2013 cohort, we preserve the overall number of reciprocal and non-reciprocal links each student has in fall 2014. We also preserve the number of reciprocal and non-reciprocal same-type links that each student has. In addition, as our results in the prior two subsections suggest that housing plays an important role in link formation, we also run simulations in which reciprocal and non-reciprocal same-house links are preserved for each student.

In practice, randomly generating graphs that satisfy all of these constraints is challenging. To simulate the constrained network, we follow the approach developed by [Fosdick et al.](#)



Figure 6: Assortativity in Homophily: Simulated and Actual



Notes: Correlation of percent same gender or ethnicity of friends. 10,000 simulated graphs via a version of the configuration method, modified to replicate the observed number of reciprocal and non-reciprocal friendships among those of same and different genders/ethnicity. The housing constraint also preserves the same number of friends by house crossed with gender/ethnicity.

(2018). We start with the observed network, and then sequentially swap links that have the same features—say, reciprocal or non-reciprocal same-gender and same-house links. We carry out 40,000 such swaps. Each simulation starts with the final network of the previous simulation. We repeat this procedure 10,000 times to generate 10,000 simulations in total.

The assortativity in homophily found in our data is unlikely to occur by chance, as illustrated in Figure 6. That figure displays the resulting correlations between the fraction of same-gender and same-ethnicity friends across linked individuals. The observed correlation is significantly higher than would be produced mechanically by chance—there is little overlap between the distribution of simulated correlations and the correlation actually observed in the data.<sup>30</sup> Nonetheless, the distributions of simulated networks for same-type friendships

<sup>30</sup>The difference between the fraction of same-gender or same-ethnicity friends or study partners does not predict friendship or study partnership retention between fall 2014 and fall 2015. As noted in Section 3.1,

are centered well above 0, reflecting the “homophily paradox” insight of [Evtushenko and Kleinberg \(2021\)](#).

Constraining the simulations to be consistent with the observed profile of friendships, both within and outside of houses, increases the simulated correlation. Intuitively, more constraints on simulated networks bring them closer to the observed network—but even with these additional constraints, the observed correlations are still statistically unlikely to occur by chance.

For mixed-gender or mixed-ethnicity friendships, the simulated correlations are centered roughly at 0, whereas the observed correlations are significantly lower than 0. That is, in the data, friendships between individuals of different types (gender or ethnicity) who both have a large fraction of same-type friendships are quite rare. Instead, different-type friendships tend to be between one person who has relatively many different-type friends, and another person who has relatively few.

## 6 Malleable Characteristics, Connections, and Outcomes

The Caltech Cohort Study elicits a variety of behavioral characteristics—such as risk attitudes, altruism, and so on—that we collectively refer to as *malleable*. In addition to being more prone to change, these characteristics are also likely more difficult to observe. In this section, we examine the dynamic patterns of homophily over malleable characteristics.

Observed similarity between linked individuals over malleable characteristics could be the outcome of either selection, assimilation, or both. This is in contrast to gender, ethnicity, or even the house individuals are associated with, which rarely change over the horizon of a friendship or study partnership.<sup>31</sup> Although similarity on malleable characteristics could also be due to selection, there is the additional possibility that interactions over time could drive similarity through a process of assimilation.

---

very few friends are retained between fall 2013 and fall 2014.

<sup>31</sup>Indeed, in our data, we only observe the declared gender and ethnic identity of students upon matriculation. Similarly, we observe only a few changes in the houses in which students reside and majors a student declares, and thus code students as belonging to the same house and major throughout their tenure.

We observe significant homophily over malleable characteristics; although it is, in general, smaller in magnitude than the homophily over gender, ethnicity, and housing, as shown in the next subsection. Of particular note is anti-homophily on GPA, which is high in the first two years, but disappears by the beginning of students’ junior year, suggesting the possibility of some assimilation on this important outcome. We then turn our attention to comparing the changes in similarity of malleable characteristics for those relationships that persist over a year, versus those that are newly formed. Changes are similar whether a relationship is new or retained from the prior year (or just a randomly chosen pair) suggesting that there is little or no assimilation on most malleable characteristics. Finally, we examine how an important outcome—GPA—changes with the gender and ethnicity of study partners. Although this analysis is exploratory, it indicates that same-gender study partnerships between females is associated with significantly improved academic performance.

## 6.1 Homophily over Malleable Characteristics

Homophily in many malleable characteristics is significant even after controlling for homophily in ethnicity, gender, and house membership, as shown in Table 4. This table reports results from a fixed-effect regression model explaining friendships and study partnerships, similar to Table 2. As in that table, the coefficients here represent the change in the number of links that are formed per 1,000 potential links, if both nodes are similar or the same on a given characteristic, controlling for similarity in other characteristics. In order to make coefficients comparable, our similarity measures for malleable characteristics are based on *median split* of responses. Namely, for each individual, we code whether each characteristic is above or below the median. For example, having similar subjective well-being means that both nodes of a potential link are either in the top-half or bottom-half of the subjective well-being distribution.<sup>32</sup>

---

<sup>32</sup>We use principal components for both the risk and altruism (dictator-giving) duplicate elicitations on the survey. Work or Sleep reflects a response to a question about the amount of sleep prior to important work obligations. See Section 2, as well as Gillen et al. 2019 and Snowberg and Yariv 2021, for further implementation details.

Table 4: Behavioral and Other Traits (per 1,000 potential links)

| Survey:                                 | Fall 2013          | Friends            |                    | Study Partners     |                    |
|---|--------------------|--------------------|--------------------|--------------------|--------------------|
|   |                    | Fall 2014          | Fall 2015          | Fall 2014          | Fall 2015          |
| Similarity: Risk Preferences            | 0.18<br>(0.25)     | 0.36<br>(0.29)     | 0.53*<br>(0.31)    | 0.80***<br>(0.27)  | 0.84***<br>(0.26)  |
| Similarity: Dictator Giving             | -0.09<br>(0.26)    | 1.75***<br>(0.31)  | 1.71***<br>(0.33)  | 1.30***<br>(0.28)  | 1.43***<br>(0.28)  |
| Similarity: IAT Gender                  | -0.18<br>(0.26)    | -0.69**<br>(0.30)  | -0.42<br>(0.32)    | -0.72**<br>(0.28)  | -0.35<br>(0.27)    |
| Similarity: IAT Race                    | 0.15<br>(0.26)     | 0.02<br>(0.30)     | -0.22<br>(0.32)    | -0.21<br>(0.28)    | 0.05<br>(0.27)     |
| Similarity: Subjective Well-Being       | -0.26<br>(0.26)    | -0.25<br>(0.30)    | 0.08<br>(0.32)     | -0.06<br>(0.28)    | 0.15<br>(0.27)     |
| Similarity: Body Mass Index             | 0.17<br>(0.26)     | 0.90***<br>(0.30)  | 0.10<br>(0.31)     | 1.10***<br>(0.27)  | 0.27<br>(0.27)     |
| Similarity: Sleep Hours                 |                    | 0.57*<br>(0.34)    | 0.40<br>(0.36)     | 0.94***<br>(0.31)  | 0.88***<br>(0.30)  |
| Similarity: Hours / Week on Video Games | -0.37<br>(0.28)    | 1.88***<br>(0.32)  | 2.11***<br>(0.33)  | 2.02***<br>(0.29)  | 1.54***<br>(0.28)  |
| Similarity: Work or Sleep               |                    | -3.98***<br>(0.64) | -3.76***<br>(0.66) | -4.96***<br>(0.59) | -3.65***<br>(0.57) |
| Similarity: Review Attendance           |                    | -1.65***<br>(0.30) | -1.11***<br>(0.31) | -1.05***<br>(0.27) | -0.73***<br>(0.26) |
| Similarity: GPA                         | -7.29***<br>(0.32) | -3.49***<br>(0.32) | -0.12<br>(0.30)    | -2.79***<br>(0.29) | 0.26<br>(0.26)     |
| Similarity: GPA Perception              | 0.57**<br>(0.25)   | 1.04***<br>(0.29)  | 1.40***<br>(0.31)  | 1.49***<br>(0.27)  | 0.57**<br>(0.26)   |
| Same Gender                             | 2.06***<br>(0.23)  | 3.45***<br>(0.27)  | 3.19***<br>(0.28)  | 2.30***<br>(0.25)  | 1.37***<br>(0.24)  |
| Same Ethnicity                          | 3.14***<br>(0.26)  | 2.90***<br>(0.29)  | 2.55***<br>(0.31)  | 2.09***<br>(0.27)  | 1.26***<br>(0.26)  |
| Same House                              | 3.14***<br>(0.34)  | 21.11***<br>(0.39) | 21.25***<br>(0.41) | 18.00***<br>(0.36) | 12.52***<br>(0.35) |
| Same Major                              | 1.25***<br>(0.32)  | 2.27***<br>(0.32)  | 3.28***<br>(0.30)  | 4.01***<br>(0.29)  | 7.72***<br>(0.26)  |
| Constant                                | 5.90***<br>(0.43)  | -0.47<br>(0.55)    | -3.54***<br>(0.56) | -1.59***<br>(0.50) | -3.49***<br>(0.47) |

Notes: \*\*\*, \*\*, \* denote statistical significance at the 1%, 5%, and 10% level with standard errors in parentheses. Similarity variables are split above and below the median for each value, with above median receiving a value of 1, and 0 otherwise. Risk is an index of risky project and qualitative risk preferences. Body Mass Index is proportional to weight divided by height.

Homophily over malleable characteristics is, in most cases, smaller in magnitude than that observed for ethnicity, gender, and housing.<sup>33</sup> Of note is the relatively strong homophily over altruism—measured in the Caltech Cohort Study through giving in the Dictator Game—that develops by the beginning of students’ sophomore year. As these regressions control for homophily over gender, ethnicity, and housing, this is not the consequence of students who are similar on one of these characteristics also being similar on malleable characteristics.

Homophily is, in general, stronger for malleable characteristics that are easier to observe, such as body mass index, generosity, how much sleep people get, and especially, the amount of time students spend playing video games.<sup>34</sup> The similarity over this final dimension increases across students’ time at Caltech, perhaps indicating an increasing commitment to their field of study, and an increased awareness of who else has chosen that course of study.<sup>35</sup> Taken together, the patterns of homophily over malleable characteristics are suggestive of a general tendency to form links with people who are similar on many different dimensions. The delayed emergence of homophily over malleable characteristics to fall 2014 indicates that it may take time to discover others’ malleable characteristics and identify those who are compatible.<sup>36</sup>

Our split of variables around the median makes coefficients comparable, and generally does not affect the trends we report. One exception is the variables pertaining to GPA. Table 4 suggests a strong anti-homophily on the important outcome of GPA, which diminishes over time. This pattern vanishes when considering raw distances between students’ GPAs as the explanatory variable, implying the GPA effects in 4 are driven by the classification of values around the median. We examine the subtle ways by which assimilation operates on GPA in the next two subsections.

---

<sup>33</sup>Results are quite similar without controls for same ethnicity, gender, and housing, dispelling the concern that collinearity between some controls and some of the malleable characteristics affects results.

<sup>34</sup>Sleep Hours refers to responses to the question “On average, how many hours of sleep do you get a night when school is in session?” Work or Sleep refers to responses to the question “Before an exam, do you go to sleep at your usual time or when you feel like your preparation work is done?”

<sup>35</sup>Some majors at Caltech are quite small. We group similar small majors, such as Geology and Geophysics.

<sup>36</sup>The lower magnitudes and significance could also be partly the result of greater noise in the observation of malleable characteristics in general, and difficult-to-observe malleable characteristics in particular.

## 6.2 Selection or Assimilation?

As noted above, changes in homophily patterns can be an artifact of both selection of new friends who exhibit greater similarities, and assimilation, whereby connected individuals become more similar to one another. We now assess whether any observed similarity of malleable characteristics is present at the start of a relationship, indicating selection as the driving force, or whether it becomes pronounced as the relationship evolves, indicating assimilation as the driving force.

There is little evidence of assimilation as relationships evolve, as shown in Table 5.<sup>37</sup> The table decomposes selection and assimilation by considering the change in distance between malleable characteristics within friendships that are retained and those that are not, as well as distance within all friendships (without the median split performed in Table 4). Specifically, we consider changes in distance between malleable characteristics, as measured in fall 2014 and fall 2015, across retained, lost, and all friends (or study partners), relative to the average change in these variables between random pairs. These changes are expressed in terms of standard deviations of the underlying variable. Differencing out the average change for random pairs is important because the distribution of each parameter in the population changes over time, in some cases increasing the average distance between individuals, and in some cases decreasing it.

While most attributes exhibit increased resemblance among links relative to the general trends observed in our sample, these differences tend to be small and insignificant. For example, retained study partners grew 0.01 standard deviations further apart in GPA than the average random link. This is directionally counter to assimilation, with a small magnitude, and statistically insignificant. Furthermore, severed friendships exhibit further separation, albeit small and statistically insignificant as well. There are a few attributes that show larger and statistically significant evidence of assimilation among retained, but not lost, friends: for example, on IAT Race. While these are of potential interest, the large number of moments

---

<sup>37</sup>This table does not examine similarity on ethnicity, gender, housing, or major. These variables are coded as fixed over time in our data, which limits our ability to identify assimilation in those.

Table 5: Changes in Distance from Friends on Behavioral Traits

|                             | Friends         |                  |                   | Study Partners     |                    |                 |
|-----------------------------|-----------------|------------------|-------------------|--------------------|--------------------|-----------------|
|                             | All             | Retained         | Lost              | All                | Retained           | Lost            |
| GPA                         | 0.02<br>(0.05)  | -0.07<br>(0.07)  | 0.07<br>(0.06)    | -0.02<br>(0.05)    | 0.01<br>(0.10)     | 0.03<br>(0.06)  |
| GPA Perception              | -0.02<br>(0.04) | -0.02<br>(0.07)  | 0.05<br>(0.06)    | 0.07<br>(0.05)     | 0.09<br>(0.10)     | -0.02<br>(0.05) |
| Risk Preferences            | 0.03<br>(0.04)  | 0.00<br>(0.06)   | -0.08<br>(0.05)   | -0.01<br>(0.05)    | -0.12<br>(0.10)    | 0.01<br>(0.05)  |
| Dictator Giving             | -0.05<br>(0.05) | -0.14*<br>(0.08) | 0.02<br>(0.07)    | -0.07<br>(0.06)    | -0.07<br>(0.12)    | 0.02<br>(0.06)  |
| IAT Gender                  | -0.05<br>(0.05) | -0.11<br>(0.08)  | -0.10*<br>(0.06)  | -0.04<br>(0.05)    | -0.18*<br>(0.11)   | -0.06<br>(0.06) |
| IAT Race                    | -0.02<br>(0.04) | -0.03<br>(0.07)  | -0.10<br>(0.06)   | -0.13**<br>(0.05)  | -0.27***<br>(0.10) | -0.06<br>(0.06) |
| Subjective Well-Being       | 0.03<br>(0.05)  | -0.00<br>(0.07)  | 0.02<br>(0.06)    | -0.00<br>(0.05)    | -0.13<br>(0.10)    | -0.01<br>(0.06) |
| Body Mass Index             | 0.04<br>(0.05)  | -0.02<br>(0.07)  | -0.12**<br>(0.05) | 0.01<br>(0.06)     | -0.09<br>(0.09)    | -0.06<br>(0.05) |
| Sleep Hours                 | 0.03<br>(0.04)  | -0.03<br>(0.07)  | -0.01<br>(0.06)   | -0.06<br>(0.05)    | -0.06<br>(0.09)    | 0.07<br>(0.06)  |
| Hours / Week on Video Games | 0.07<br>(0.06)  | 0.09<br>(0.09)   | 0.13<br>(0.08)    | 0.06<br>(0.06)     | 0.02<br>(0.14)     | 0.07<br>(0.07)  |
| Work or Sleep               | -0.05<br>(0.05) | -0.00<br>(0.08)  | -0.04<br>(0.07)   | -0.19***<br>(0.06) | -0.19*<br>(0.11)   | 0.05<br>(0.07)  |
| Review Attendance           | 0.06<br>(0.04)  | 0.03<br>(0.07)   | 0.06<br>(0.06)    | 0.06<br>(0.05)     | 0.07<br>(0.09)     | 0.04<br>(0.05)  |

Notes: \*\*\*, \*\*, \* Coefficients are changes from fall 2014 to fall 2015 in standard deviations of the underlying variable. All refers to anyone named a friend in the corresponding survey. Retained refers only to the subset that were named friends in both surveys. Lost refers to the subset that were named friends in Fall 2014, but not Fall 2015. Average distance is the average over all dyads, not just those that are connected.

we examine raises the possibility that these few statistically significant results are spurious: in particular, 8/72 of the coefficients in Table 5 are significant at the  $p < 0.1$  level.<sup>38</sup> Table 5 also indicates that lost friends are not people who have moved further apart in behaviors.

Together, the patterns in Table 4 and 5 suggest that homophily on malleable characteristics is maintained—and sometimes strengthened—by choosing relationships with individuals who are already similar. As noted above, one place where homophily strengthens—from anti-homophily to no homophily—is GPA. In the next subsection, we conduct a more detailed analysis of GPA patterns.

### 6.3 Outcomes: GPA and Homophily

A large literature examines whether homophilous environments improve academic performance, especially in the context of single-sex education for women (see the survey in Robinson et al., 2021). In this subsection, we provide an exploratory analysis of the association between changes in students' GPAs and homophilous friendships and study partnerships.

The largest association between peer attributes and changes in a student's GPA appears to stem from same-gender female study partners, as shown in Table 6. The first three columns of this table regress the change in a student's GPA between the 2015-2016 academic year and the 2014-2015 academic year on the characteristics of that student's declared study partners in the fall of 2015.<sup>39</sup> The first column reveals a slight negative association between same-ethnicity study partners and change in GPA: every additional same ethnicity study partner reduces GPA by an average of 0.06 points. The second column shows that the apparent null effect of same-gender study partnerships masks the fact that, for women, these partnerships have a significantly positive association with improved GPAs.<sup>40,41</sup>

---

<sup>38</sup>Zhang and King (2021) document a network of physicians. They find that differences in contentious prescribing lead ties to weaken or dissolve altogether, but do not affect tie formation. In contrast, for the attributes we consider, we do not see a pronounced effect of dissimilarity leading to dissolution of connections.

<sup>39</sup>Focusing on grade changes allows us to reduce concerns that the reported results are due to selection of friends or study partners on the basis of grades. Nonetheless, our results should not be interpreted as causal.

<sup>40</sup>These results are in line with the conclusions of Cools et al. (2019), who observe distinct responses in academic performance among grade-school girls and boys. See also the references therein.

<sup>41</sup>When considering interactions with ethnicity, our data suggest the most pronounced positive effect is derived for student partners who are both Asian women. Regressing the change in a student's GPA between



The positive association between GPA change and females having same-gender study partners is not an artifact of simply studying with people with higher GPAs, as shown in the third column of Table 6. This column separates the effects of partners who exhibit weaker performance, where the GPA difference is negative (denoted as “Below Partner”), from partners who exhibit stronger performance, where the GPA difference is positive (denoted as “Above Partner”). Thus, “GPA Difference in 2014-2015 AY from Below Partner (negative)” is negative if the partner’s GPA was below the student in question, and 0 otherwise, while “GPA Difference in 2014-2015 AY from Above Partner (positive)” is positive if the partner’s GPA was above the student in question, and 0 otherwise. The third column indicates that having a study partner who has a higher GPA is associated with an increase in the student’s own GPA, while having a study partner with a lower GPA has little or no association with changes in a student’s own GPA. More importantly, however, the coefficient on same-gender female study partnerships is relatively unchanged, while the coefficient on same-ethnicity study partnerships becomes insignificant.

The effects described in the preceding two paragraphs are largely absent for friends who are not study partners, as shown in the final three columns. While homophily does not seem to play an important role in how friendships relate with students’ GPA, friends do still have an impact. In particular, in the last column, we see that students’ GPAs move towards that of their friends who are not study partners, regardless of whether those friend have higher or lower GPAs. As the magnitudes of these two coefficients are similar, these associations tend to cancel out in the aggregate, leaving students’ GPA unchanged on net by their friends who are not study partners. This pattern differs from that observed within study partnerships, in column 3, where the aggregate effect of links are significant and positive.

These results highlight the importance of considering different types of relationships—here, friendships and study partnerships. Focusing on one type alone, as frequently done, paints an incomplete—here, inaccurate—picture of interactions’ impact on outcomes.

---

the 2014-2015 academic year and the 2013-2014 academic year on the characteristics of that student’s declared study partners in the fall of 2014, results weaken substantially. This merits further investigation, but could be driven by the fact that students’ classes in the first two freshman quarters are graded on a pass or fail basis. The 2013-2014 academic year GPA is, therefore, a relatively noisy measure.

Table 6: GPA Change Associated with Gender and Ethnicity of Links

| Dependent Variable:  | 2015-2016 AY GPA minus 2014-2015 AY GPA |                    |                    |                                     |                  |                    |
|--|---|--------------------|--------------------|-------------------------------------|------------------|--------------------|
|  | Study Partners                          |                    |                    | Friends that are not Study Partners |                  |                    |
| Same Gender  | 0.04<br>(0.029)                         | 0.01<br>(0.032)    | 0.00<br>(0.030)    | -0.01<br>(0.035)                    | -0.02<br>(0.039) | -0.02<br>(0.038)   |
| Same Gender $\times$ Female                                  |   | 0.13***<br>(0.046) | 0.14***<br>(0.043) |                                     | 0.05<br>(0.045)  | 0.05<br>(0.042)    |
| Same Ethnicity   | -0.06*<br>(0.033)                       | -0.06*<br>(0.032)  | -0.04<br>(0.029)   | -0.02<br>(0.030)                    | -0.02<br>(0.030) | -0.01<br>(0.028)   |
| Same House   | 0.02<br>(0.026)                         | 0.02<br>(0.026)    | 0.02<br>(0.024)    | 0.04<br>(0.037)                     | 0.04<br>(0.037)  | 0.04<br>(0.034)    |
| Same Major   | 0.05<br>(0.035)                         | 0.04<br>(0.034)    | 0.04<br>(0.033)    | -0.04<br>(0.027)                    | -0.04<br>(0.027) | -0.04<br>(0.026)   |
| GPA Difference in 2014-2015 AY<br>From Below Peer (negative) |   |                    | 0.06<br>(0.041)    |                                     |                  | 0.19***<br>(0.047) |
| GPA Difference in 2014-2015 AY<br>From Above Peer (positive) |   |                    | 0.32***<br>(0.071) |                                     |                  | 0.17**<br>(0.086)  |
| Constant   | 0.00<br>(0.039)                         | 0.00<br>(0.039)    | -0.04<br>(0.040)   | 0.03<br>(0.059)                     | 0.03<br>(0.059)  | 0.04<br>(0.054)    |

Notes: \*\*\*, \*\*, \* denote statistical significance at the 1%, 5%, and 10% level with standard errors, clustered at the participant level, in parentheses.

## 7 Conclusion

We document the extent of homophily in the composition and evolution of friendships and study partnerships using comprehensive longitudinal data on Caltech undergraduates. Our analysis produces multiple stylized facts on homophily as networks form and evolve, using a broad, incentivized survey and institutional data on students' demographics, housing, and academic performance.

First, homophily on gender and ethnicity is stable over time. Second, homophily manifests in similar ways across network layers—friendships, study partnerships, and housing. Third, homophilous tendencies on gender and ethnicity are persistent at the individual level, suggesting that homophily is a stable trait. Fourth, we see homophily over homophily: more homophilous individuals tend to connect with more homophilous others. Fifth, there is little evidence of assimilation on malleable characteristics: homophily appears to be driven predominantly by the selection of connections. Sixth and finally, homophily has important, but nuanced, impacts on outcomes. It operates differently through study partnerships and friendships. In particular, ethnic homophily in study partnerships slightly reduces students' GPA gains, while gender homophily leads to GPA gains, but only for females.

Given the non-trivial evolution of networks in our data, our study highlights the importance of tracking networks over time. Our study also highlights the importance of tracking multiple layers of interactions in the population—in our case, friendships, study partnerships, and housing—as well as multiple dimensions over which homophilous tendencies might be present. The effects of network layers vary substantially: for instance, study partners and friends have different impacts on students' GPAs. Furthermore, homophily across dimensions exhibits non-trivial interactions.

Our results suggest that engineering social and academic connections can be a challenging task, regardless of a designer's objective. In our setting, while students tend to connect with others in their house and in their major, those tendencies do not significantly alter the homophilous features of their friends and study partners. In particular, on-campus residential arrangements may have a limited effect on the diversity of students' connections, in line with

the findings of [Carrell et al. \(2013\)](#) and [Fosnacht et al. \(2020\)](#). Nonetheless, more research is needed to ascertain the dynamics of interactions in very small groups—in the college setting, these can be clubs, athletic teams, and the like. Such groups can limit substantially the profile of potential connections, and offer an instrument for altering the profile of members' connections. Alternatively, if the available connections within such small groups are not to members' liking, members may instead seek connections outside their group.



- de Klepper, Maurits, Ed Sleenbos, Gerhard van de Bunt, and Filip Agneessens**, “Similarity in friendship networks: Selection or influence? The effect of constraining contexts and non-visible individual attributes,” *Social Networks*, 2010, *32*, 82–90.
- De La Haye, Kayla, Harold D. Green Jr., David P. Kennedy, Michael S. Pollard, and Joan S. Tucker**, “Selection and influence mechanisms associated with marijuana initiation and use in adolescent friendship networks,” *Journal of Research on Adolescence*, 2013, *23* (3), 474–486.
- Evtushenko, Anna and Jon Kleinberg**, “The paradox of second-order homophily in networks,” *Scientific Reports*, 2021, *11* (1), 1–10.
- Fosdick, Bailey K., Daniel B. Larremore, Joel Nishimura, and Johan Ugander**, “Configuring random graph models with fixed degree sequences,” *Siam Review*, 2018, *60* (2), 315–355.
- Fosnacht, Kevin, Robert M. Gonyea, and Polly A. Graham**, “The relationship of first-year residence hall roommate assignment policy with interactional diversity and perceptions of the campus environment,” *The Journal of Higher Education*, 2020, *91* (5), 781–804.
- Fu, Feng, Martin A. Nowak, Nicholas A. Christakis, and James H. Fowler**, “The evolution of homophily,” *Scientific Reports*, 2012, *2* (1), 1–6.
- Gale, Douglas and Lloyd S. Shapley**, “College Admissions and the Stability of Marriage,” *The American Mathematical Monthly*, 1962, *69* (1), 9–15.
- Garlick, Robert**, “Academic peer effects with different group assignment rules: Residential tracking versus random assignment,” *American Economic Journal: Applied Economics*, 2018, *10* (3), 345–369.
- Gillen, Ben, Erik Snowberg, and Leeat Yariv**, “Experimenting with measurement error: Techniques with applications to the caltech cohort study,” *Journal of Political Economy*, 2019, *127* (4), 1826–1863.
- Goeree, Jacob K., Margaret A. McConnell, Tiffany Mitchell, Tracey Tromp, and Leeat Yariv**, “The 1/d law of giving,” *American Economic Journal: Microeconomics*, 2010, *2* (1), 183–203.
- Graham, Bryan S.**, “Homophily and transitivity in dynamic network formation,” *mimeo*, 2016.
- Heckman, James J. and Stefano Mosso**, “The economics of human development and social mobility,” *mimeo*, 2014.
- Jackson, Matthew O.**, “A survey of network formation models: Stability and efficiency,” *Group Formation in Economics: Networks, Clubs, and Coalitions*, 2005, *664*, 11–49.
- , *Social and Economic Networks*, Princeton University Press, 2010.
- , “Inequality’s economic and social roots: The role of social networks and homophily,” *mimeo*, 2021.
- , **Brian W. Rogers, and Yves Zenou**, “The economic consequences of social-network structure,” *Journal of Economic Literature*, 2017, *55* (1), 49–95.
- Kandel, Denise B.**, “Homophily, selection, and socialization in adolescent friendships,” *American Journal of Sociology*, 1978, *84* (2), 427–436.
- Mathews, Heather and Alexander Volfovsky**, “Latent community adaptive network regression,” *mimeo*, 2021.

- McPherson, Miller, Lynn Smith-Lovin, and Craig Rawlings**, “The enormous flock of homophily researchers: Assessing and promoting a research agenda,” *Personal Networks: Classic Readings and New Directions in Egocentric Analysis*, 2021, p. 459.
- , –, and **James M. Cook**, “Birds of a feather: Homophily in social networks,” *Annual Review of Sociology*, 2001, *27* (1), 415–444.
- Nahemow, Lucille and M. Powell Lawton**, “Similarity and propinquity in friendship formation,” *Journal of Personality and Social Psychology*, 1975, *32* (2), 205.
- Osgood, D. Wayne, Daniel T. Ragan, Lacey Wallace, Scott D. Gest, Mark E. Feinberg, and James Moody**, “Peers and the emergence of alcohol use: Influence and selection processes in adolescent friendship networks,” *Journal of Research on Adolescence*, 2013, *23* (3), 500–512.
- Overgoor, Jan, Bogdan State, and Lada A. Adamic**, “The structure of U.S. college networks on Facebook,” in “Proceedings of the International AAAI Conference on Web and Social Media,” Vol. 14 2020, pp. 499–510.
- Pearson, Michael, Christian Steglich, and Tom Snijders**, “Homophily and assimilation among sport-active adolescent substance users,” *Connections*, 2006, *27* (1), 47–63.
- Powell, G. Bingham**, “Political cleavage structure, cross-pressure processes, and partisanship: An empirical test of the theory,” *American Journal of Political Science*, 1976, pp. 1–23.
- Ready, Elspeth and Eleanor A. Power**, “Measuring reciprocity: Double sampling, concordance, and network construction,” *Network Science*, 2021.
- Robinson, Daniel B., Jennifer Mitton, Greg Hadley, and Meagan Kettley**, “Single-sex education in the 21st century: A 20-year scoping review of the literature,” *Teaching and Teacher Education*, 2021, *106*, 103462.
- Sacerdote, Bruce**, “Peer effects with random assignment: Results for Dartmouth roommates,” *The Quarterly Journal of Economics*, 2001, *116* (2), 681–704.
- Shrum, Wesley, Neil H. Cheek Jr., and Sandra MacD**, “Friendship in school: Gender and racial homophily,” *Sociology of Education*, 1988, pp. 227–239.
- Small, Mario L. and Laura Adler**, “The role of space in the formation of social ties,” *Annual Review of Sociology*, 2019, *45*, 111–132.
- Snijders, Tom**, “Stochastic actor-oriented models for network dynamics,” *Annual Review of Statistics and its Application*, 2017, *4*, 343–363.
- Snowberg, Erik and Leeat Yariv**, “Testing the waters: Behavior across participant pools,” *American Economic Review*, 2021, *111* (2), 687–719.
- Somashekhar, Mahesh H.**, “Diversity through homophily? The paradox of how increasing similarities between recruiters and recruits can make an organization more diverse,” *McGill Sociological Review*, 2014, *4*, 1–18.
- Song, Yangbo and Mihaela van der Schaar**, “Dynamic network formation with incomplete information,” *Economic Theory*, 2015, *59* (2), 301–331.
- Vaquera, Elizabeth and Grace Kao**, “Do you like me as much as I like you? Friendship reciprocity and its effects on school outcomes among adolescents,” *Social Science Research*, 2008, *37* (1), 55–72.
- Veenstra, René, Jan Kornelis Dijkstra, Christian Steglich, and Maarten H.W. Van Zalk**, “Network–behavior dynamics,” *Journal of Research on Adolescence*, 2013, *23* (3), 399–412.

- Verbrugge, Lois M.**, “The structure of adult friendship choices,” *Social Forces*, 1977, 56 (2), 576–597.
- Zafarani, Reza, Mohammad Ali Abbasi, and Huan Liu**, *Social Media Mining: An Introduction*, Cambridge University Press, 2014.
- Zhang, Victoria and Marissa D. King**, “Tie decay and dissolution: Contentious prescribing practices in the prescription Drug epidemic,” *Organization Science*, 2021, 32 (5), 1149–1173.
- Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov**, “Political effects of the internet and social media,” *Annual Review of Economics*, 2020, 12, 415–438.
- Zuckerman, David**, “Unseen preferences: Homophily in friendship networks,” *mimeo*, 2022.



# A Additional Analyses

Figure A.1: Correlations between percentages of same gender/ethnicity of new and old study partners with no types, stochastic block types, and fixed types.

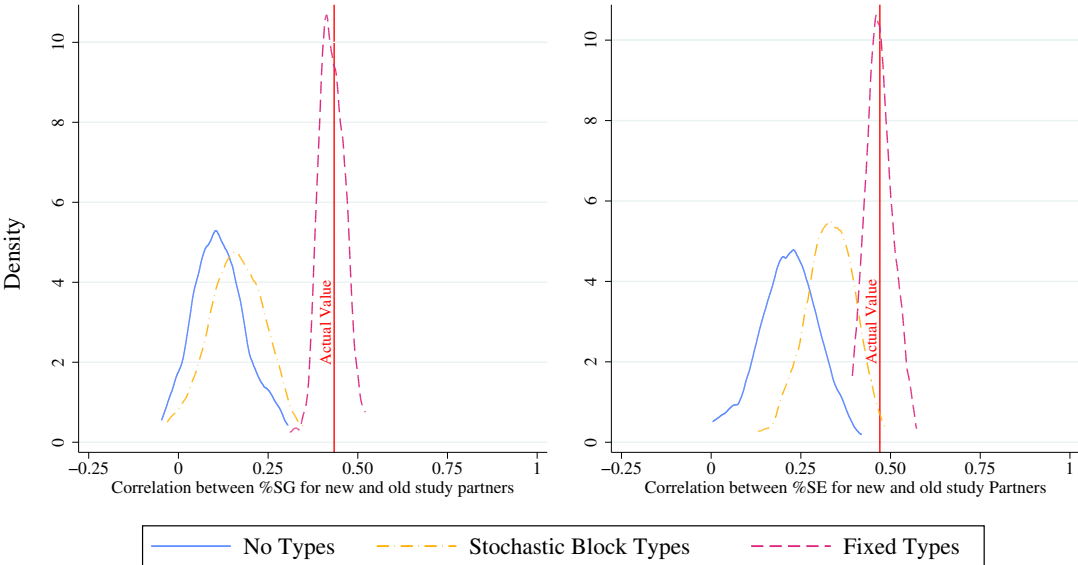


Table A.2: Regression of Links on Types of Dyad, Scaled per 1,000 Potential Links

|                                     | Friends            |                    | Study Partners     |                    |
|-------------------------------------|--------------------|--------------------|--------------------|--------------------|
| Panel A: Freshman Year (Fall 2013)  |                    |                    |                    |                    |
| Same Gender                         | 2.07***<br>(0.233) | 1.44***<br>(0.276) |                    |                    |
| Same Ethnicity                      |                    | 3.12***<br>(0.259) | 2.00***<br>(0.372) |                    |
| Same Gender<br>x Same Ethnicity     |                    |                    | 2.07***<br>(0.497) |                    |
| Constant<br>(Non-Homophilous)       | 1.31***<br>(0.168) | 1.52***<br>(0.135) | 0.74***<br>(0.201) |                    |
| Panel B: Sophomore Year (Fall 2014) |                    |                    |                    |                    |
| Same Gender                         | 3.47***<br>(0.271) | 2.55***<br>(0.325) | 2.37***<br>(0.248) | 1.87***<br>(0.299) |
| Same Ethnicity                      |                    | 3.42***<br>(0.296) | 1.86***<br>(0.425) | 2.64***<br>(0.271) |
| Same Gender<br>x Same Ethnicity     |                    |                    | 2.92***<br>(0.569) | 1.59***<br>(0.522) |
| Constant<br>(Non-Homophilous)       | 2.20***<br>(0.196) | 3.01***<br>(0.159) | 1.64***<br>(0.237) | 2.60***<br>(0.146) |
| Panel C: Junior Year (Fall 2015)    |                    |                    |                    |                    |
| Same Gender                         | 3.34***<br>(0.280) | 2.28***<br>(0.338) | 1.59***<br>(0.239) | 1.01***<br>(0.288) |
| Same Ethnicity                      |                    | 3.18***<br>(0.312) | 1.42***<br>(0.443) | 0.80***<br>(0.377) |
| Same Gender<br>x Same Ethnicity     |                    |                    | 3.33***<br>(0.593) | 1.76***<br>(0.266) |
| Constant<br>(Non-Homophilous)       | 2.35***<br>(0.202) | 3.12***<br>(0.168) | 1.92***<br>(0.246) | 1.81***<br>(0.505) |
|                                     |                    |                    | 2.14***<br>(0.172) | 1.90***<br>(0.209) |

Notes: \*\*\*, \*\*, \* denote statistical significance at the 1%, 5%, and 10% level with standard errors in parenthesis.

## B Screenshots

This section contains screenshots for the questions that are analyzed in this paper. For a sample of screenshots of an entire survey,

see [lyariv.mycpanel.princeton.edu/~papers/ScreenshotsFall2014.pdf](http://lyariv.mycpanel.princeton.edu/~papers/ScreenshotsFall2014.pdf).

Figure B.1: Friendships

Here, we'd like you to tell us which other Caltech students you socialize and study with. Please keep in mind that all names specified on this survey are automatically converted into anonymized numerical IDs, so that there is no way for anyone to observe what you fill out here.


Please name the five (5) other Caltech students you socialize with the most by choosing their names from the drop-down lists that will appear when you start typing their names. Also, please tell us the average number of hours you spend with each of these friends on a weekly basis (a number between 0 and 100, with up to one decimal point). If you spend time will less than five other students on a weekly basis, it is fine to leave the last few lines blank.

Friend 1:   
Average number of hours you spend together per week:   
Friend 2:   
Average number of hours you spend together per week:   
Friend 3:   
Average number of hours you spend together per week:   
Friend 4:   
Average number of hours you spend together per week:   
Friend 5:   
Average number of hours you spend together per week:

On average, how many hours a week do you spend socializing with friends?

Submit

Figure B.2: Study Partners



## Caltech Cohort Study

Please name the five (5) other Caltech students you study with the most. If you study with less than five other students on a weekly basis, it is fine to leave the last few lines blank.

Study Partner 1:


Study Partner 2:

Study Partner 3:

Study Partner 4:

Study Partner 5:

Figure B.3: GPA Perceptions



## Caltech Cohort Study

We'd like to know where you think you will rank, in your class, in GPA, for the upcoming year. That is, where do you think, in percentile terms, your GPA for *only classes you will take this year* will compare with the GPAs of your classmates?

Enter an integer number between 1 and 99, where 99 means you think you'll be in the *top percentile*, and 1 means you think you'll be in the *bottom 1 percent*:

Figure B.4: Risk Aversion: Project



## Caltech Cohort Study

You are endowed with 200 tokens (or \$2) that you can choose to keep or invest in a risky project. Tokens that are not invested in the risky project are yours to keep.

The risky project has a 40% chance of success.

- If the project is successful, you will receive 3 times the amount you chose to invest.
- If the project is unsuccessful, you will lose the amount invested.


Please choose how many tokens you want to invest in the risky project. Note that you can pick any number between 0 and 200, including 0 or 200:

95

You will learn your payoff in this section at the end of the survey.

Submit

Figure B.5: Risk Aversion: Qualitative, Work / Sleep, and Sleep Hours



## Caltech Cohort Study

How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?  
Please tick a box on the scale, where the value 0 means: 'not at all willing to take risks' and the value 10 means: 'very willing to take risks'

0  1  2  3  4  5  6  7  8  9  10

---

On average, how many hours of sleep do you get a night when school is in session?  
 Hours


---

Before an exam, do you go to sleep at your usual time or when you feel like your preparation work is done?

I go to sleep at my usual time  
 I go to sleep when my work is done

---

Figure B.6: Dictator Game



## Caltech Cohort Study

You now have **300 tokens to be divided between you and another**, randomly chosen, survey participant.

All other survey participants will be given the same choice: that is, they will be given 300 tokens to divide between themselves and another participant.

Your payoff from this section will be how much you allocate to yourself, plus how much is allocated to you by another randomly chosen participant. Note that the recipient, the participant that receives money from you, and the participant that you receive money from will be different, and both will be chosen randomly.

Amount for you:  Amount for recipient:

(Amounts entered should be numbers between 0 and 300.)

Figure B.7: IAT Race, Example 1



**Caltech Cohort Study**

**African American**  
**Positive Word**

**European American**  
**Negative Word**

**Press Space Bar to Begin**

Keep your index fingers on the 'e' and 'i' keys

Figure B.8: IAT Race, Example 2



**Caltech Cohort Study**

**African American**  
**Positive Word**


**European American**  
**Negative Word**



**Bad**

Keep your index fingers on the 'e' and 'i' keys

Figure B.9: IAT Gender, Example



# Caltech Cohort Study


**Male**  
**Science**

**Female**  
**Liberal Arts**

**Physics**

Keep your index fingers on the 'e' and 'i' keys

Figure B.10: Subjective Well Being (current)



# Caltech Cohort Study

Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?

**Submit**



Figure B.11: Subjective Well Being (future)



## Caltech Cohort Study

Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. Just your best guess, on which step do you think you will stand in the future, say about five years from now?

Submit

Figure B.12: BMI (height and weight)



## Caltech Cohort Study

What is your height?  feet  inches

What is your weight in pounds?  lbs

Figure B.13: Videogames

Over the past month, how many hours a week would you estimate you have spent playing video games?

 hours

Figure B.14: Review Attendance



## Caltech Cohort Study

How frequently do you attend review sessions?

Please tick a box on the scale, where the value 0 means: 'I never attend review sessions' and the value 10 means: 'I never miss a review session'

0  1  2  3  4  5  6  7  8  9  10