The Determinants and Consequences of Friendship Composition[•]

Jason M. Fletcher
Yale University

Stephen L. Ross
University of Connecticut

Yuxiu Zhang
Yale University

This Version: April 16, 2013

Abstract

This paper examines the demographic pattern of friendship links among youth and the impact of those patterns on own educational outcomes using the friendship network data in the Add Health. We develop and estimate a reduced form matching model to predict friendship link formation and identify the parameters based on across-cohort, within school variation in the "supply" of potential friends. Our model provides novel evidence on the impact of small changes in peer demographic composition on the pattern of friendship links suggesting, for example, that increases in the number of students from college educated family backgrounds leads to a greater likelihood of friendship links with students of that type among students with either mothers who are college educated and high school graduates and that increases in the share of African-American or Hispanic students leads to reductions in the incidence of cross race friendships. We then use the predicted friendship links from the model in an instrumental variable analysis of the effects of friends' socioeconomic status as measured by parental education on own grade point average outcomes. Although the conditional correlation between friendship composition and grade point average suggests large associations between friends' characteristics and own grades, this effect is robust only for females in the instrumental variable analysis. We then present evidence that the GPA effects are driven by science and English grades and a mechanism is likely through self-esteem.

[•] The authors thank the NICHD (1R21HD066230) for its financial support. The authors thank Edward Vytlacil, Ken Frank, and participants of the 2012 Add Health Users Conference for helpful suggestions.
This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

Introduction

A growing body of evidence has documented the effects of peers on the academic outcomes of school children.[1] The effect of peers on such outcomes raises the natural policy question of what would really happen if peers were changed. A change in peer composition may alter the social dynamics of a school or other social network, and as a result alter the social ties and personal interactions through which peer effects may operate. Several recent studies (Fletcher and Ross 2012, Calvó-Armengol et al. 2009) document the effect of friendship and social networks on student and youth outcomes. Weinberg (2007) shows that students tend to associate with individuals like themselves, which may mitigate the impact of any change in peer composition. Mayer and Puller (2008) show evidence that increasing the opportunities of heterophilous (e.g. cross-race) relationships is not sufficient to substantially increase these links. Finally, Carrell, Sacerdote and West (2011) conduct a policy experiment where Air Force Academy students are assigned to work groups intended to maximize the performance of the lowest ability students. They find that in their treatment group, students sort into subgroups based on ability, eliminating the positive peer effects identified in earlier studies of the same environment, indicating the need for further information on how friendships form before policies can be suggested. However, the difficulties with estimating the friendship matching process in "real world" settings are numerous. This paper combines a quasi-experimental research design within the setting of high school friendship network formation to extend the literatures on friendship formation, as well as estimating the consequences of friend composition on educational outcomes.

In particular, this study uses the friendship nomination data in the Add Health sample to study the causal impact of small changes in peer composition on the demographic pattern of friendship formation.[2] We focus on within grade (or cohort) friendships, which represent a supermajority (83% of same-sex nominations[3]) of friendship ties in our sample, and identify the effect of peer composition on friendship formation by exploiting across cohort and within school variation in the composition of students (i.e. "potential friends"). Specifically, we classify potential friendship ties within grade based on the demographic match between each pair of students, and examine heterogeneity in the effects of cohort demographic composition on the likelihood of friendship formation, conditional on demographic 'type' of student pair by school fixed effects. Balancing tests confirm that cohort composition is orthogonal to incidental student attributes within school-pair-type cells. Across cohort variation is regularly exploited in studies of the effects of peers on student outcomes (beginning with Hoxby 2000), but to our knowledge this is the first study to exploit this variation in order to examine friendship formation.[4] Our

[1] See Lavy and Schlosser (2011), Bifulco, Fletcher and Ross (2011) and references contained.

[2] Few other national datasets contains information on nominated best friends. Additionally many datasets contain a single grade-level (cohort) from each sampled school (e.g. NELS, ECLS-K, ELS, etc).

[3] In this paper, our focus is within school friendship. Among all within school same-sex friend nominations in which both parties have identifiable student id, school id and grade id, 17% are cross-grade nominations. Among all same-sex nominations with identifiable friend id (including those with missing school/grade id), 66% are same-grade within school; 14% are cross-grade within school; for another 19% nominations, nominated friends' id is identifiable, but school id and grade id are not, therefore we don't know whether they are within school/grade or not; for the rest 1%, we know the two parties in the nominations are from two different identifiable schools. In terms of out-of-school nomination, we need to take account unidentifiable nominations coded as 77777777, which indicate out-of-school nomination. The proportion of out-of-school nomination is about 16% and the proportion of uncertain 14%.

[4] Perhaps the closest papers to our study in this regard are those by Fisman and colleagues (2008), who used random assignment during speed dating interactions to estimate the preferences for same and opposite-race social ties (i.e.

model of friendship formation focuses on peer maternal education based on the importance of parental education for child outcomes (Haveman and Wolfe 1995) and a previous finding in the same sample that peer's maternal education has a significant impact of academic outcomes (Bifulco, Fletcher and Ross 2011). We also examine race, ethnicity and gender, given the well-known concentration of friendships among students of the same race, ethnic and gender (Moody 2001). Student race/ethnicity is especially important to investigate given the large race/ethnic differences in educational attainment in the population.

Specifically, we examine how differences in the socio-demographic composition of a cohort relative to the composition of the other cohorts in the same school affect the likelihood of any pair of same gender students to mutually identify each other as friends. On maternal education, our key sorting results are for females, and we find that as the number of students whose mothers have a college degree increases relative to mothers with high school degrees, friendships between female students whose mothers both have a college degree or where one mother has a college degree and the other completed high school become more likely. Our estimates imply that a 10 percentage point increase in the number of maternal college students is on average associated with a 7 percent and 10 percent increase in the probability of forming links between two maternal college students and between a maternal college and a maternal high school graduate student, respectively. Given the focus on friendship link formation, these results cannot be driven simply by an increase in the opportunity for college educated friends, but instead are consistent with an increase in the attractiveness of maternal college educated friends as the number of maternal college educated students rises. On race/ethnicity, differences in the share black have the largest impact, with white-white friendships becoming more frequent for both men and women and Other-Other (predominantly Asian-Asian) friendships becoming more frequent for women. Hispanic-Other friendships also become less likely among women. Increases in share Hispanic lead to lower rates of female across race friendships (white-Hispanic and Hispanic-Other) and higher rates of black-black friendships among men. Therefore, increases in minority representation appear to increase the level of homophily in friendship formation, consistent with Mayer and Puller (2008) among others.[5]

We next examine the effect of friendship patterns on student outcomes. Building on earlier work (Bifulco, Fletcher and Ross 2011), we focus our analysis on the impact of maternal education levels on academic outcomes. As discussed by Manski (1993, 2000) and others, research of the effects of social interactions between individuals must address several empirical issues because individuals select into friendships and peer groups. In order to address these concerns, we use our estimated model of the formation of friendship links in order to develop predictions of (i.e. instruments for) friendship composition for individual students. These instruments are highly predictive of individual student's actual friendship patterns even though the predictions do not contain any information on the individual's friendship patterns and are only identified by across-cohort variation in the demographic composition of schools. We find that the number of friends with a college educated mother has a large positive effect on the grade point average of female students, where a one standard deviation increase in the number of maternal college friends is associated with a 0.165 standard deviation increase in GPA. These

---

dates). As in that study, our study examines the effects of presumably exogenous changes in the opportunity set for forming interpersonal relationships

[5] The difficulty of producing heterophilous ties in groups when adding diversity is also a likely mechanism for why the experiment conducted by Carrell et al. (2012) that increased academic diversity among military squadrons reduced the outcomes of these individuals.

effects operate primarily through higher grades in English and Science courses. Mechanism analyses suggest that the GPA effects may be driven in part by self-esteem effects, consistent with the "role model effect" discussed by Durlauf (2004). The mechanism analysis also shows that the majority of the effect of maternal college friends operates through an effect on students whose mothers have a college education for both GPA and the mechanism variables. While there is a strong conditional correlation between the proportion of friends with a college educated mother and grade point average for male students, these effects do not persist in our instrumental variable estimates. The effect of the number of high school drop-out friends is zero in both the OLS and IV estimates for both male and female students, indicating asymmetric effects of maternal schooling.

In terms of identification, the strength of the GPA effects of friends arises from the exogeneity of the instrument. The student level GPA regressions include school by student demographic type fixed effects. Further, we address concerns about incidental parameters bias in the fixed effect estimates by calculating individual specific friendship predictions using individual specific fixed effects that omit any information associated with that individual's friendship choices. Following Guryan, Kroft and Notowidigdo (2009), we address the bias caused by omitting this information on the individual's choices by developing a control function for inclusion in the GPA regressions, and balancing tests confirm that the resulting instruments are not correlated with predetermined attributes conditional on the fixed effects and the control function. If there is a weakness of our identification strategy, it relates to our exclusion restrictions. While we rule out general cohort level peer effects by including grade by school fixed effects, we cannot rule out the possibility that a randomly assigned peer environment that leads to more friendships with students whose mothers have a college education for a given type of student also directly increases girls' GPA. A somewhat weaker conclusion based on our results is that peer environments that raise the likelihood of a particular student having friends with college educated mothers lead to an increase in girls' GPA.

Finally, for girls, we conduct a series of calculations and simulations examining the effect of an increase in the number of students with a mother who is a college graduate. The calculations examine the direct effect of adding more maternal college students both through the increase in the opportunity to form such friendships and the estimated effect of the share maternal college on the likelihood of friendship formation. A ten percentage point increase in the share maternal college in each cohort is associated with a 45 percent increase in the number of friends for the maternal college subsample, 107 percent increase for the maternal high school graduate subsample, and 111 percent increase for the maternal drop-out subsample. Most of these changes are associated with the increase in opportunities for friendships with maternal college students as opposed to the 9 and 14 percent effects of changing the probability of friendship formation. Our simulations allow for the effect of changes in the racial composition of students as the number of maternal college students is increased. Over two different scenarios, these simulations indicate substantially smaller increases in the number of maternal college friends for both students who mothers have a college degree (a 20 percent increase) and for students who mothers are a high school graduate (approximately a 60 percent increase). The calculated effect for the maternal drop-out subsample is relatively stable as we allow for effects of changes in racial composition.

Empirical Model of Friendship Formation

Consider a sample of schools (s) with a set of grades or cohorts (c) in each school. Students of a given gender may be systematically allocated to a school through their parents' choices, but are assumed to be distributed randomly across the cohorts or grades in any school because parents cannot easily observe the composition of individual cohorts when choosing a school, especially when those grade compositions will only be determined at a future time.[6]

Within a grade, every student can potentially form a friendship with any other student in the grade, and our student friendship data can be rearranged as a sample of pairs of students i and j where students are categorized into one of *m* nominal "types" where student *i* is type *x* and student *j* is type *y*. The establishment of a social link between any pair of students (P_{ijcs}) may be described by the following linear probability model

$$P_{ijcs} = \tau_{cs}\left(\beta_{xy}Z_{cs} + \delta_{xys}D_{ij}^{xy} + \varepsilon_{is}\left(1+\gamma_y Z_{cs}\right) + \varepsilon_{js}(1+\gamma_x Z_{cs}) + \theta\varepsilon_{is}\varepsilon_{js} + \mu_{ijcs}\right) \quad (1)$$

where Z_{cs} is a 1 by m vector measuring the demographic composition of each cohort over types, β_{xy} captures our behavior of interest by allowing the likelihood of friendship formation for each pair of student types {x,y} to vary with the demographic composition of the cohort, τ_{cs} allows the probability of friendship formation overall to vary across cohorts so that probabilities of formation with a particular individual can fall with cohort size, D_{ij}^{xy} is a 1 by m(m-1)/2 dimension vector of dummy variables (i.e. fixed effects) indicating whether an individual pair represents a match of individuals of types x and y, δ_{xys} allows the effect of belonging to that pair type {x,y} on friendship formation to vary by school so that the estimates of β_{xy} are identified by across cohort comparisons of friendship patterns within school and friendship type, student unobservables on the propensity to form friendships are captured by random effects ε_{is} and ε_{js}, the effect of this propensity is allowed to vary by cohort composition and with the propensity of the person to which the individual is matched since such heterogeneous responses to cohort composition might bias estimates of the direct effect of cohort composition on friendship formation, and finally μ_{ijcs} is a stochastic return to the match between these particular students.

Taking the conditional expectation of equation (1), yields

$$E\left[P_{ijcs}|\tau_{cs},Z_{cs},D_{ij}^{xy}\right] =$$
$$\tau_{cs}\begin{pmatrix}\beta_{xy}Z_{cs} + \delta_{xys}D_{ij}^{xy} + E\left[\varepsilon_{is}|\tau_{cs},Z_{cs},D_{ij}^{xy}\right]\left(1+\gamma_y Z_{cs}\right) + \\ E\left[\varepsilon_{is}|\tau_{cs},Z_{cs},D_{ij}^{xy}\right](1+\gamma_x Z_{cs}) + \theta E\left[\varepsilon_{is}\varepsilon_{js}|\tau_{cs},Z_{cs},D_{ij}^{xy}\right] + E\left[\mu_{ijcs}|\tau_{cs},Z_{cs},D_{ij}^{xy}\right]\end{pmatrix} \quad (2)$$

and illustrates the required assumptions for consistent estimates of β_{xy} and δ_{xys}

$$E\left[\varepsilon_{is}|\tau_{cs},Z_{cs},D_{ij}^{xy}\right] = 0$$
$$E\left[\varepsilon_{is}\varepsilon_{js}|\tau_{cs},Z_{cs},D_{ij}^{xy}\right] = 0$$
$$E\left[\mu_{ijcs}|\tau_{cs},Z_{cs},D_{ij}^{xy}\right] = 0 \quad (3)$$

[6] This assumption is supported in our sample by balancing tests conducted in Bifulco, Fletcher and Ross (2011) and later in this paper, demonstrating that individual attributes of students are not correlated with within-school variation in cohort composition.

We believe that the assumptions in equation (3) are reasonable given our earlier assumption of the random allocation of students of each type x to a particular cohort c within a school s, and the construction of the sample to include all possible pairs of students in a grade. Our and the literature's concern about bias arises from the potential correlation between ε_{is} and school composition (Z_{cs}) based on students (or their parents) sorting systematically into schools based on the demographic composition of those schools potentially violating the first condition in equation (3). Further, this sorting likely varies with the students' demographic attributes so that the conditional distribution of ε_{is} within school is not constant across students of different types. However, by linearly conditioning on school (s) by student pair type (x,y) fixed effects, we condition out the effect of sorting into schools on the mean of the distribution of ε_{is} for each observable student type and, given quasi-random assignment to cohorts within schools, ε_{is} should be uncorrelated with the within school variation in cohort demographics.[7]

Given the first assumption in equation (3), the only possible mechanism for violating the second assumption is if ε_{is} and ε_{js} are correlated within school and cohort. Specifically,

$$\begin{aligned} &E\left[\varepsilon_{is}\varepsilon_{js}|\tau_{cs}, Z_{cs}, D_{ij}^{xy}\right] = \\ &\quad Cov\left[\varepsilon_{is}, \varepsilon_{js}|\tau_{cs}, Z_{cs}, D_{ij}^{xy}\right] + E\left[\varepsilon_{is}|\tau_{cs}, Z_{cs}, D_{ij}^{xy}\right]E\left[\varepsilon_{js}|\tau_{cs}, Z_{cs}, D_{ij}^{xy}\right] \\ &\quad = Cov\left[\varepsilon_{is}, \varepsilon_{js}|\tau_{cs}, Z_{cs}, D_{ij}^{xy}\right] \end{aligned} \qquad (4)$$

However, our sample of pairs within cohort are constructed to include all possible pairs of students and so with the assumption of no selection into cohorts within schools the correlation or covariance must be zero. Finally, it is relatively standard to assume that the idiosyncratic error associated with the match between two individuals μ_{ijcs} is orthogonal to the observables.[8]

In the context of our specific problem and data, we next specify the details of the model that we will estimate. First, we note that asymptotically τ_{cs} must be inversely proportional to the number of potential friends in cohort (n_{cs}) because otherwise the actual number of friends will limit to either 0 or infinity as the cohort size becomes larger. As a result we approximate τ_{cs} with $1/n_{cs}$ and estimate δ_{xys} and β_{xy} using the following equation

$$P_{ijcs} = \beta_{xy}\left(\frac{Z_{cs}}{n_{cs}}\right) + \tilde{\delta}_{xys}D_{ij}^{xy} + \tilde{\varepsilon}_{is}\left(1 + \gamma_y Z_{cs}\right) + \tilde{\varepsilon}_{js}(1 + \gamma_x Z_{cs}) + \theta\tilde{\varepsilon}_{is}\tilde{\varepsilon}_{js} + \tilde{\mu}_{ijcs} \qquad (5)$$

[7] The above claim relies on the implicit assumption that the expectation of ε_{is} conditional on type x is zero otherwise random variation in cohort racial composition will lead to systematic changes in the average unobservables of the individuals in a type and cohort. However, this restriction is a standard assumption in virtually all reduced form studies including studies that exploit random assignment because one cannot randomly assign the attributes of the randomly assigned factors, e.g. peers or environmental circumstances, and our analysis captures the causal effect of more students of a given type in a cohort on friendship formation including the effect through unobservables that are systematically associated with that type.

[8] In principle, one might question whether students have correlated unobservables in the same cohort because some of them will end up in the same classroom or share similar interests. However, such phenomena do not lead to a conditional correlation within the population unless that likelihood varies systematically across cohorts in the same school. The effect of the average probability of sharing a class or an interest with another student on friendship link formation should be captured by the school-student pair type fixed effects, and after conditioning out that effect the only obvious source of correlation is sorting, which our assumptions rule out.

Note that, at least to a first order approximation, the pair-type by school fixed effects ($\tilde{\delta}_{xys}$) can be estimated as a common set of parameters across cohorts within a school, $\tilde{\delta}_{xys} = \frac{\delta_{xys}}{\bar{n}_{cs}} \approx \frac{\delta_{xys}}{n_{cs}}$, because with a moderate size or larger school and quasi-random allocation of students to cohorts n_{cs} is relatively constant within a school (near the mean - $\bar{n}_{cs}$) and deviations in n_{cs} within school can be treated as exogenous.[9]

Data and Estimation of the Friendship Model

Data Description

In order to examine the determinants and achievement consequences of friendship ties during high school, we use the only available dataset with information on nominated friends from multiple grade-levels in a large number of schools, the National Longitudinal Study of Adolescent Health (Add Health). Add Health is a school based longitudinal study of health and education-related behaviors of adolescents with follow up through age 30. For this paper, we focus on the "In-School" data collection, which utilized a self-administered survey to more than 90,000 students in grades 7 through 12 during a class period at school between September 1994 and April 1995. The survey focused on collecting data on socio-demographic characteristics, family background, health status, risk behaviors, academic achievement, school factors, and friendship nominations. Specifically, each student respondent was asked to identify up to five male and five female friends that attended the same school (these nominations were later cross-referenced with school rosters). Based on the friendship nominations, social networks within each school can be constructed, allowing data links between friends' reported background characteristics and respondent's reported course grades in English, math, science, and history courses.

Of the over 90,000 students originally surveyed, there are several sample size reductions necessary to create our analysis sample. 178 individuals were dropped from the sample due to missing identification numbers[10]; another 2,666 are dropped because of missing grade, race, sex, mom's education, or missing the majority of their friendship information; we exclude 112 observations from small schools (less than 40 students in school or less than an average of 10 students per grade); we exclude the twin sample, which contains 2508 students. This process gives us an empirical sample of 84,654 coming from 139 schools with school size between 44 and 2,367 students allocated across two to four grades or cohorts.

Like much previous work, we focus on same-gender friendships in our analysis. The primary reason for this choice is to separate "friends" from "romantic relationships". We also limit our analysis to examining links between individuals in the same grade level. As we

[9] One concern with equation (1) arises from the heteroscedasticity associated with the linear probability model. With similar number of friends at the individual level regardless of cohort size, the matrix of P_{ijcs} becomes very sparse for large schools with large numbers of students in each cohort and is much more dense for smaller schools. Equation (5) addresses this by decreasing the magnitude of the independent variable for large cohorts/schools where the frequencies of non-zero P_{ijcs} are very low rather than requiring the effect of cohort composition to be the same in percentage point terms for link formation in allowing for a lower probability of link frequency for these sparse regions of the social link vector.

[10] These individuals were likely new students and not yet on the school roster

describe in more detail below, this focus allows us to utilize an across-cohort research design[11]. We focus on directed ties, and mutual friendship in particular, meaning two students are considered as a pair of friends if they both nominated each other. We assume the influence from friends to be strongest in a relationship which both parties in the pair agree on the friendship. It is also worth noticing that in Add Health, though a student can nominate up to five same-gender friends, not many students appear constrained by this cap. The average numbers of identifiable same-gender friends nominated are 2.61 for male students and 3.07 for female students. The majority of the nominations are one-direction. Therefore, the number of mutual ties is low. On average, a male has 0.68 and a female has 1.09 mutual friends.[12] We begin by showing the basic friendship patterns in the data on our key variables of interest. Table 1 shows the fraction of same-gender/same-grade friendships in each maternal education category by the maternal education of the student. The rows identify the type of student being considered and the columns identify the type of friends, with panel 1 presenting the average and percentages for females and panel 2 for males. The bottom row shows the population shares of each group. The table is consistent with substantial homophily in friendship patterns over maternal education through the combined effect of sorting into schools and sorting into friendship. Looking along the diagonal of each panel, the percent of friends with the same maternal education as the student always exceeds the fraction of students in the population of that type. Females appear to exhibit higher levels of homophily than males at lower levels of maternal education.

Table 2 shows the same patterns by race and ethnicity. Again, the table is consistent with even higher levels of homophily since the fraction of own race friends far exceeding the fraction of that race in the population. Black and Hispanic females exhibit higher levels of homophily than black and Hispanic males. In general, Table 1 and 2 also indicate that females have more friends than males, and students with college graduate maternal education have more friends than others. In order to separate the effect of school level segregation and homophily within schools, we also present the deviation of friendship frequencies within individual schools from expected friendship frequencies based on school level demographic composition. Appendix Table 1A and 2A confirm substantial homophily by maternal education and racial/ethnic groups within school.

Evidence Supporting the Research Design

To provide evidence that our use of across-cohort, within school variation is valid and uncontaminated by other unobservables, we conduct a series of balancing tests (following Bifulco et al. 2011, Lavy and Schlosser 2011, Billings et al. 2012) that estimate the associations between the cohort measures and individual-level exogenous attributes, such as age, health status, nativity status, etc. In Table 3, we regress cohort composition over maternal education, race and ethnicity on ten exogenous attributes of students, omitting the student themselves from

[11] Although the focus on same-grade nominations may appear constricting, we note that over 80% of all nominations we capture in the data are for individuals in the same grade. We also show in Table 3 that our cohort variables (i.e. the 'supply' of types of friends in a cohort) is unrelated to whether individuals nominate friends outside of their grade.

[12] We also examine link models based on assuming a friendship exists when there is a link between the pair in at least one direction. The resulting estimates on the effect of demographic composition on link formation are very similar to the result presented here.

this composition.[13] The specific cohort variables we use in both these balancing tests and our friendship formation model are percentage of black students, Hispanic students, students whose mom graduated from college, and students whose mom dropped out from high school, by grade-gender within schools. Each column in Table 3 represents a single regression of relating cohort composition on variables of individual characteristics of interest along with controls with school-gender fixed effects and cohort fixed effects.[14] Our results are consistent with cohort characteristics of interest that are conditionally plausibly exogenous (within schools) in that they cannot be explained by the predetermined attributes of the students in the cohort. Of the 40 individual t-tests, one is significant, and for the four regressions, none of the F-tests on the set of 10 variables is significant.

Estimating the Matching Model

Next, we describe the construction of our matched sample. For each student, we form a pair between him/her and each of the rest of the students from the same grade and gender. This process results in a fully matched sample of potential links in every school-cohort-gender cell. The size of the matched sample is about 12 million directed links, or 6 million unique pairs. For our friendship formation model, the outcome is a binary variable indicating whether the two parties in a pair nominated each other as their friend.

We defined four racial and ethnic categories (non-Hispanic white, non-Hispanic black, Hispanic, and Asians/Other race[15]) and 4 maternal education categories (four year college degree, high school graduate/some college, high school drop-out, and maternal education not reported). This implies 10 unique racial/ethnic combinations and 10 unique maternal education combinations of the two parties in a pair. Further, race/ethnicity and maternal education together define 16 student types. This results in 136 potential student-pair combinations for pair type-school fixed effects.[16] Finally, in order to obtain a parsimonious vector β_{xy} we restrict the interactions of pair type with cohort demographic composition so that cohort maternal education composition only affects friendship formation through the maternal education attributes of the pair of students, and, similarly, cohort racial and ethnic composition is restricted to only operate through the racial and ethnic attributes of the pair. That is, we do not allow interactions between the race types of the pair and cohort measures of maternal education levels.

[13] These balancing tests follow Billings, Deming and Rockoff (2012) reversing the regression relationship, as compared to Bifulco et al. 2011 and Lavy and Schlosser 2011, and placing the cohort composition on the left hand side so that a single F-test can be used to examine whether the set of exogenous attributes can systematically explain the within school by type variation associated with each cohort composition variable. Following Guryan et al. (2009) the balancing test models also control for school level composition omitting the student's contribution in order to address the mechanical negative correlation between student's own attributes and cohort composition variables that omit the student. However, our cohorts are sufficiently large that the balancing tests results are very similar whether or not the Guryan et al. control is included in the models.

[14] It is important to point out that the chance of a student nominating friends out of his/her own grade is not correlated with any of the cohort variables, suggesting that cross-grade friendship is not impacted by cohort composition.

[15] The majority of this group are Asian (70.46% indicate themselves not White, Black, Hispanic, Native American or other (not Asian). 56.06% clearly identify themselves as Asian), and results are robust to omitting non-Asians from this group. In some context below, we refer the "other" racial group as Asian when "other" may cause confusion.

[16] N(N+1)/2=4(4+1)/2=10; N(N+1)/2=(16*17)/2=136. An example of a pair type is white-dropout/white-college, indicating that one party of the pair is white with a high school dropout mom, and the other party of the pair is white with a college graduate mom.

Specifically, we estimate the effects of cohort composition on the likelihood of "types" of friendship pairs forming:

$$P_{ijcs} = \beta_{race}\left(\frac{Z_{scg}^{race}}{n_{scg}}\right)D_{ij}^{Race} + \beta_{mom}\left(\frac{Z_{scg}^{mom}}{n_{scg}}\right)D_{ij}^{Mom} + \tilde{\delta}_{xysg}D_{ij}^{xy} * S * G + \tilde{\mu}_{ijcs} \qquad (5')$$

Where P_{ijcs} is the probability of a two way link between ego *i* with alter *j* and is a function of a large set of indicators variables reflecting the school and the potential pairs' type and interactions between type and cohort-school composition in the type (for example, pair type for race is interacted with the proportion of black cohortmates). Z_{scg}^{race} is the percentage of black and percentage of Hispanic in a school-cohort-gender group (with s for school, c for cohort/grade and g for gender); Z_{scg}^{mom} is the percentage of college graduate maternal education and high school dropout maternal education; n_{scg} is the number of students in a school-cohort-gender cell. D_{ij}^{Race} and D_{ij}^{Mom} are dummies of pair type based on race and maternal education respectively. $D_{ij}^{xy} * S * G$ represents the interaction of race-education combined pair type and school by gender. The large set of fixed effects constrain comparisons between individuals of the same pair type who attend the same school but are in different grade levels (cohorts) and are thus exposed to different cohort compositions. Because of the clear difference between male and female in Table 1-2, we estimate this model by gender subsamples in order to present coefficients separately.[17]

Empirical Results—Matching Model

Table 4 presents the results related to the maternal education status of grade-mates. The estimate in each cell is the coefficient from the interaction of a certain pair type dummy and cohort variables of maternal education. The main finding is that, for females, increases in grademates with college educated mothers increases the likelihood of college grad/college grad pairs being friends as well as the likelihood of college grad/high school grad pairs. For a sense of the magnitude of these effects, a 10 percentage point increase in the proportion of grademates with college educated mothers among females would increase the likelihood of a college grad/college grad pair being friends by 6.8% (about 0.09 percentage point relative to 1.3% baseline for an average cohort-gender cell[18]), and also increase the likelihood of a college grad/high school grad friendship link by 10.2% (0.09 percentage point relative to 0.9% baseline).[19] Note that the results are relatively noisy and uninformative for the coefficients on percent maternal drop-out for friendships involving students whose mothers have a college degree due to the negative correlation in share maternal college and maternal drop-out across schools.

[17] As a check of our model, in Appendix Table 3A we show that our predicted number of mutual friends in total and by demographic categories are very close to the actual numbers at mean level.

[18] Recall that the matching model estimates the likelihood of all potential same-grade/same-gender matches, so that the sample size for the full sample is 12 million potential pairs. Thus, the likelihood that any individual pair are friends is small. See Appendix Table 4A

[19] To give an example, for a grade with an average size of 100 girls, if the percentage of students with college educated mothers increases by 10 percentage points, which means the scaled cohort variable for college educated mothers increases by 0.001, the likelihood of a college grad-college grad pair to be friends will increase by 0.001242. Among all college grad-college grad ties, the proportion of actually formed mutual friends is 0.013 (see Table 4A). It means college grad-college grad ties increases by 9.6%(=0.001242/0.013). With the actual average cohort size of 142, the increase is about 6.8%(=9.6%*100/142) or 0.09 percentage point increase(=0.001242*100/142).

These results cannot be driven simply by an increase in the opportunity for college educated friends because we are estimating the probability of a specific link being formed. Rather, these results are consistent with an increase in the attractiveness of maternal college educated friends as the number of maternal college educated students rises, possibly because, given homophily, more individuals with college educated mothers leads to individuals with college educated mothers being more socially connected and therefore generating a greater social return associated with such friendships (Ballester et al. 2006). It is important to note that similar magnitude and same sign results exist for these two pair variables interacted with percent high school drop-out mothers, but those are much less precisely estimated, potentially due to the small size of this group in the population. As a result, our key findings should be interpreted as the effect of an increase in the number of students with college graduate mothers in a cohort.

In Table 5 we present estimates of our matching model for the likelihood of various "types" of friendship links based on same-race or different-race matches. We find that increases in the share of blacks and Hispanics at the grade-level appear to increase homophily and decrease heterophily in friendship formation along specific dimensions. We find for females that increases in the proportion of black students in the grade increase same-race friendships for white and Other students (increases homophily) and reduces different-race friendships in white/Other and Hispanic/Other potential pairs (reduces heterophily). These effects are also non-trivial—a 10% increase in the proportion of black students in a grade increases the likelihood that a white/white pair is formed by 10.6% or 0.13 percentage point off a base of 1.2% actual mutual friends formed. [20]

Similarly, for males we find that an increase in the proportion of black students in the grade also increases the likelihood of same race links, for white and Hispanic pairs. Increases in the proportion of Hispanic grade-mates imply reductions in the likelihood of different-race pairs for Black/Other and White/Hispanic for females (reduced heterophily) and an increase in the likelihood of same-race pairs for black males (increased homophily).

The main results from our matching model suggest that exogenous changes in the composition of class/schoolmates leads to changes in the likelihoods of the "types" of friend-pairs found in the data. More specifically, the results present direct evidence of increases in preferences for homophily relative to heterophily in this sample, especially with regards to race and ethnicity. While the literature has consistently found evidence of homophily, we know of no other work that documents this shift towards homophily as the population of minority groups increases. Further, in our analysis, the shift is identified using a quasi-random research design to estimate effects so that these changes cannot be attributed to other school level environmental changes that might often accompany equilibrium changes in demographic composition.

Estimating the Effect of Friendship Composition on Academic Outcomes

Model Outline

[20] Using a similar example to the earlier footnote on education, for a grade with an average size of 100 girls, if the percentage of black students increases by 10%, which means the scaled cohort variable for black increases by 0.001, the likelihood of a white-white pair to be friends will increase by 0.001795. Among all female white-white ties, the proportion of actually formed mutual friends is 0.012 (See Table 4A in Appendix). It means white-white ties increases by 15.0%(=0.001795/0.012). With the actual average cohort size of 142, the increase is about 10.6%(=15.0%*100/142) or 0.13 percentage point increase(=0.001795*100/142).

Using our estimated model of friendship formation, we next develop predictions of friendship composition for individuals of any specific type in a specific cohort and school. An individual's predicted friendship outcome in terms of number of friends can be expressed based on summing the expression in equation (5') over all matches within the cohort[21]

$$p_{ics} = \sum_{j\neq i, j\in\{c,s\}}(\beta_{xy}\left(\frac{Z_{cs}}{n_{cs}}\right) + \tilde{\delta}_{xys}D_{ij}^{xy} + \theta\tilde{\varepsilon}_{is}\tilde{\varepsilon}_{js} + \tilde{\mu}_{ijcs}) + \tilde{\varepsilon}_{is}\sum_{j\neq i, j\in\{c,s\}}(1+\gamma_y Z_{cs})$$

$$+(1+\gamma_x Z_{cs})\sum_{j\neq i, j\in\{c,s\}}\tilde{\varepsilon}_{js} \quad (6)$$

By dropping terms involving the unobservables, we define the deterministic component of friendship outcomes as

$$\bar{p}_{xcs} = \sum_{j\neq i, j\in\{c,s\}}\left(\beta_{xy}\left(\frac{Z_{cs}}{n_{cs}}\right) + \tilde{\delta}_{xys}D_{ij}^{xy}\right) \quad (7)$$

for any i of type x, since the deterministic component does not vary across individuals of the same type, school and cohort.

Similarly, using our model parameter estimates, we define the predicted friendship outcomes as

$$\hat{p}_{xcs} = \sum_{j\neq i, j\in\{c,s\}}\left(\hat{\delta}_{xys}D_{ij}^{xy} + \hat{\beta}_{xy}\left(\frac{Z_{cs}}{n_{cs}}\right)\right) \quad (8)$$

where $\hat{\delta}_{xys}$ and $\hat{\beta}_{xy}$ are based on the model in equation (4). Again, the predictions do not vary across individuals of the same type, school and cohort.

A key problem that arises from the estimation of $\hat{p}_{xcs}$ is that the estimates only vary across cohorts, $\hat{p}_{xcs} \neq \hat{p}_{xds}$ where $c \neq d$, if the total number of students of type *x* in school *s* is not large; otherwise cohort composition will simply represent school composition ($Z_{cs} \approx Z_{ds}$ for all cohorts *c* and *d* in a school). In fact, the estimates of $\hat{\beta}_{xy}$ on which $\hat{p}_{xcs}$ are based is only identified because Z_{cs} varies across cohorts. Therefore, while the total number of students of type *x* in the sample and the total number of students in any school or cohort may be relatively large, the number of students in each type in each school must be relatively small in order to create variation across cohorts. While our estimates of $\hat{\beta}_{xy}$ are consistent in the number of schools under the assumption of a linear probability model and the assumptions in equation (2), the dimensionality of our fixed effect vector increases linearly with the number of schools and the number of pair types, and so the school by student pair-type fixed effects suffer from an incidental parameters bias due to small numbers of observations in each cell. Specifically, the unobservable of a student i of type x in school s affects the estimates of $\hat{\delta}_{xys}$ for all types y and so the conditional expectation of the unobservable in the friendship choice equation $\tilde{\varepsilon}_{is}$ is non-zero.

$$E\left[\tilde{\varepsilon}_{is}\middle|\hat{\beta}_{xy}\left(\frac{Z_{cs}}{n_{cs}}\right), \hat{\delta}_{xys}D_{ij}^{xy}\ \forall\ y\right] \neq 0 \quad (9)$$

[21] Note that the predicted number of friends of a given race, ethnicity or maternal education can be found by summing equation (4) over all matches within the cohort with students in that demographic category.

In order to address this source of bias, we develop an individual specific measure of predicted friendship outcomes that explicitly omits all pairs involving individual i from fixed effects associated with pairs involving individuals of type x. First, in a linear probability model, consistent estimates of $\hat{\beta}_{xy}$ can be and were obtained above by simply differencing out the school by pair type fixed effects in equation (5') and estimating

$$(P_{ijcs} - \bar{P}_{xys}) = \beta_{xy}\left(\frac{Z_{cs}}{n_{cs}} - \left(\frac{\overline{Z_{cs}}}{n_{cs}}\right)_s\right) + (\widetilde{\omega}_{ijcs} - \bar{\omega}_{cs}) \quad (10)$$

where $\widetilde{\omega}_{ijcs}$ is the sum of all terms involving unobservables in equation (5') and the bar operator implies the mean of the preceding term over all observations in a school by pair-type cell.

In mean differenced models, the standard approach to estimating the fixed effects is to back out those fixed effects by calculating the mean of the within cell residual in the non-differenced sample. The individual specific fixed effect that omit pairs involving the individual i in cohort c can be estimated in the same way by summing the predicted residual over all cohorts d and pairs of students, k and j, with at least one student of type x other than student i.

$$\hat{\delta}_{xys}^{-i} = \sum_{d\in\{s\}}\left(\sum_{k\neq i\ if\ d=c,k\in\{x,d,s\}}\sum_{j\neq k\ \&\ j\neq i\ if\ d=c,j\in\{d,s\}}\left(P_{xycs} - \hat{\beta}_{xy}\left(\frac{Z_{cs}}{n_{cs}}\right)\right)\right) \quad (11)$$

The notation in equation (11) is structured so that the first summation term sums over all cohorts in the school in order to calculate a school level fixed effect, the second term sums over all other students in the same cohort and of same type x as student i, and then the third term sums over all students of type y in the same cohort excepting students i and k if types x and y are the same.

Now based on equation (8), we define the individual specific prediction as

$$\hat{p}_{ics}^{-i} = \sum_{j\neq i, j\in\{c,s\}}\left(\hat{\delta}_{xys}^{-i}D_{ij}^{xy} + \hat{\beta}_{xy}\left(\frac{Z_{cs}}{n_{cs}}\right)\right) \quad (12)$$

However, as noted by Guryan et al. (2009), this process creates a negative correlation within type-cohort-school because an individual's contribution to the fixed effect is eliminated for themselves and not for anyone else in the type-cohort-school.[22] Guryan et al. proposes a solution to this bias for peer composition or subgroup means which is to include an additional control for peer composition at a higher level of aggregation, also omitting self. This control captures the negative correlation arising from omitting self and the estimates on the subgroup means are unbiased.[23]

[22] In Guryan et al.' example, players select into golf tournaments, but are then randomly assigned to teams, which Guryan refers to as urns. The average team ability experienced by an individual golfer (omitting self) is negatively correlated (conditional on tournament fixed effects) with the individual's unobservable because within the tournament and urn the golfer cannot be paired with him/herself.

[23] In order to apply the Guryan et al. logic to our example, it is useful to consider a slight generalization to their problem. Consider the following simple behavioral model

$y_{ics} = \beta X_{cs} + \delta_s + \pi_{ics}$

where *c* is an urn and *s* is a tournament. Assume that for any individual *i*, X_{cs} is correlated with ε_{ics} , but can be divided into two additively separable components

In our context, students of a given type x sort into schools, but their allocation to a cohort or grade is assumed to be quasi-random. Therefore, the aggregate groups (or tournaments) are defined as type-school cells, and type-cohort-school cells are equivalent to one of Guryan et al.'s subgroups (or urns). We wish to separate the predicted friendship outcome from equation (8) into a component that omits all information involving choices made by individual i and a second component that contains this contamination.

$$\hat{p}_{ics} = \hat{p}_{ics}^{-i} + \hat{q}_{ics} \tag{13}$$

The expression $\hat{p}_{ics}^{-i}$ has been constructed in equation (12) so that it does not contain any information on the unobservable of individual i, and differencing equations (8) and (12) yields

$$\hat{q}_{ics} = -\tau_{cs} \sum_{j \neq i, j \in \{c,s\}} (\hat{\delta}_{xys} - \hat{\delta}_{xys}^{-i}) D_{ij}^{xy} \tag{14}$$

For our context, this contaminated component is equivalent to the control developed by Guryan et al. because it contains the contributions of the individual's choices to the conditional mean that is represented by the fixed effect estimates. The inclusion of this control will eliminate the bias caused by omitted an individual's own contribution to the fixed effect estimates in constructing predicted numbers of friends.

Finally, consider an empirical model of an outcome y_{ics} where a student of type *x*'s outcome may be influenced by the type of social links formed by the student:

$$y_{ics} = \theta \underline{p}_{ics} + \gamma_{xs} + \tau_{ics} \tag{15}$$

where $\underline{p}_{ics}$ is a vector of friendship composition outcomes, such as number of friends and number of friends of different demographic groups, γ_{xs} is a vector of school by student type fixed effects, and p_{ics} potentially correlates with the unobservable τ_{ics}.

Therefore, we estimate a series of first stage models where the friendship composition outcome depends upon the individual level prediction of composition, a second term containing the contaminated component of the prediction, and the school by type fixed effects.

$X_{cs} = X_{ics}^{i} + X_{ics}^{-i}$
where the first component contains the contamination that leads to the correlation and the second component is uncorrelated with ε_{ics}
$E\left[\pi_{ics} \middle| X_{ics}^{i}, \delta_s\right] = \alpha X_{ics}^{i}$

$E\left[\pi_{ics} \middle| X_{ics}^{-i}, \delta_s\right] = 0$
The second component X_{ics}^{-i} is equivalent to the average urn ability omitting self, and simply including this control will lead to biased estimates because X_{ics}^{i} is omitted and X_{ics}^{i} and X_{ics}^{-i} are correlated. However, as suggested by Guryan et al., including both variables yields unbiased estimates since
$E[y_{ics} | X_{ics}^{-i}, X_{ics}^{i}, \delta_s] = \beta X_{ics}^{-i} + (\beta + \alpha) X_{ics}^{i} + \delta_s$
While the Guryan et al. idea of controlling for the tournament mean minus the individual's contribution seems intuitively appealing, the true source of the solution is that the within tournament variation in this mean nearly perfectly correlates with the individual, additively separable portion of the mean (the contaminated component) that has been removed from the variable of interest.

$$p_{ics} = \omega_1 \hat{p}_{ics}^{-i} + \omega_2 \hat{q}_{ics}^{-i} + \varphi_{xs} + \tilde{\rho}_{ics} \quad (16)$$

where any element of p_{ics} depends only on the same elements of $\hat{p}_{ics}^{-i}$ and $\hat{q}_{ics}^{-i}$, e.g. number of friends or number of friends whose mothers have a college education, so that the coefficients in equation (17) are scalers.

We propose to obtain consistent estimates of θ in equation (15) using a second stage estimation equation based on the estimates of equation (16) as follows

$$y_{ics} = \theta \underline{\hat{p}}_{ics} + \pi \underline{\hat{q}}_{ics}^{-i} + \tilde{\gamma}_{xs} + \tilde{\tau}_{ics} \quad (17)$$

where this equation also includes the predicted composition based on equation (16) and the contaminated component of the instrument in order to avoid the Guryan et al. bias.

Data Description

In the following two sections, we first present the descriptive statistics of the student level data relevant to our examination of friendship effects of GPA. Then we present our estimates. After describing the data, we begin the empirical analysis by presenting standard OLS models that links the GPA of friends together. However, these models are likely biased due to endogeneity of friends. We next show that, using our matching model from the previous section, we can predict the "types" of friends that individuals nominate in the data using across-cohort variation in the "supply of friend-types". We then incorporate our predicted friendship patterns as instruments in a two-stage analysis to examine the importance of endogeneity. As we show above in the context of our matching model, we also present balancing test results that show that individual covariates are unrelated to our instruments in Appendix 7A

Table 6 shows the distribution of maternal education and means of GPA by racial/ethnic groups at student level. The proportion of high school dropout is the lowest for White, and Hispanic students have significantly lower maternal education than the other three groups. For all racial/ethnic groups, female students have higher average GPA than males; black and Hispanic students show lower GPA than the other two groups. We also provide pooled descriptive statistics for the key variables used in our analyses in Appendix Table 5. Among students in our sample, 91% were born in the U.S., 92% report living with their mother and the average family size is 4.3 persons per household. The average age is 15, and 40% of sample come from grade 9 and 10.

Empirical Results—Effects of Friends on Academic Achievement

Table 7 presents estimates of the effects of friend composition of maternal education on students' GPA for female and male sub samples. Each column represents results from a single regression with school-type FE and school-grade FE. The OLS coefficients from column (1) and (4) shows students having more friends with a college educated mother have higher GPA relative to their grade mates, but having more friends whose mom dropped out from high school doesn't significantly correlate with lower GPA. The pattern shows no gender difference.

Our next step is to leverage the predicted friendship pattern measures we extract from the matching model above to use as instruments for actual friendship patterns. Like any instrument, our measures need to be strongly related to the endogenous (actual) friendship pattern and unrelated to the unobservables determining GPA. In Appendix Table 6A, we show that our predicted friendship composition measures are strongly related to the actual friendship nominations in the data, where the F-statistics are between 50-260, even after controlling for school by type fixed effects and eliminating any effect of individual's own friendship choices. As we show in Appendix Table 7A our instruments are largely unrelated to a large set of observable factors ("balancing tests"), which is consistent with the exclusion restriction.

In Columns 2, 3, 5, 6 in Table 7, we then examine friend composition effects for academic achievement using two-stage least squares. Considering the low number of mutual friends on average, we test one "type" of friend at a time. For example, in column (2), we regress GPA on the actual number of friends with a college graduate mom, instrumented by the predicted number of friends whose moms graduated from college.[24] The first observation is that the IV estimates differ from OLS estimates. The coefficient of college graduate mom increases by 30% for female, but changes from significant and positive to insignificant, small and negative for males. The coefficient of dropout from high school remains relatively small and statistically insignificant. As noted above, the F-stat from first stage of the 2SLS is in the range of 30-250, by which we can reject the null hypothesis of weak instruments.

The estimated coefficient of peer college graduate mom is 0.196, indicating that one more mutual friend with a college educated mom is associated with a 0.20 grade point increase of GPA, which is about a 6.8 percent increase at a mean GPA of 2.89 for all female students and represents a 0.257 standard deviation increase in GPA. Multiplying by the standard deviation of number of maternal college friends, the effect of a one standard deviation in the number of friends of this type is a 0.165 standard deviation increase in GPA. In contrast, the number of friends with mom dropped out from high school is not significantly correlated with GPA. Further, for males, maternal education of friends does not show any significant effect on own GPA in the IV regression, in contrast to the OLS regression. The small and insignificant IV coefficient for males must be interpreted with some caution because the IV standard errors for the effect of having friends whose mothers are college graduates is substantially larger for the male than the female sample.[25]

Mechanisms

[24] Relevant Guryan type controls are always included in both first and second stages.

[25] To further examine the robustness of our main results, we examine the sensitivity of the results to the specification of the IV model, and whether other aspects of friendship composition might directly influence academic outcomes and simply be correlated with maternal education of friends. In Appendix Table 8A, we present results from a set of IV models. First, instead of testing one instrument at a time as in previous tables, we explore whether including both friends with high maternal education and low maternal education influences our results. Then we also examine the effect of controlling for the total number of friends and the racial and ethnic composition of friends. The positive effect of friends with high maternal education on own GPA for female is robust through the three specifications we test (column 1-3). Still, no distinguishable impact of friends' maternal education is found for males. Adding the total number of friends does not change the pattern of correlation between friends composition and own GPA for either females or males. None of the coefficients of number of black or Hispanic friends is significantly different from zero, suggesting that the effect from racial composition of friends is quite weak when controlling for friends' maternal education.

Next, in order to further examine the overall GPA effects for females, Table 8 decomposes the result based on the four subject areas of grades available in the data (math, English, science, and history). The evidence suggests that the gain in GPA from having a friend with a highly educated mother is based on better performance in both English and Science classes, but not Math and History Courses. Mirroring the main results, we find no effects for males.

In order to investigate the potential channel through which girls are affected by close friends with high maternal education, we use the preferred IV specification to test a series of outcomes of self reported behavior, beliefs, and physical and mental health status. Given our interest in identifying consistent patterns of results, we also indicate findings that are significant at the 10 percent level for this analysis. To reduce the number of tests, we use factor analysis procedures to classify variables into seven categories—self evaluation,[26] judgment regarding social environment,[27] mental status, trouble in school activities, misbehavior, smoking and drinking, and self reported health. A high score reflects high self evaluation, comfortable social environment, good mental health, having more trouble at school, more misbehavior, high frequency of smoking/drinking and good physical health respectively. More details of the factor analysis are in Appendix Table 9A.

In Table 9, we present the results of our mechanisms analyses. Each column of Table 9 refers to a single outcome of interest, and each cell represents the relevant coefficient of interest from a separate IV regression. The results suggest that female's subjective evaluation regarding self and school are consistently positively correlated with the number of friends with high maternal education they have. Female students with more friends of high maternal education are more confident and comfortable with themselves and the people around them. The results also indicate that girls with more friends whose moms graduate from college are less likely to display depression symptoms or misbehave/act out in school. The findings for self-evaluation and social environment are most notable because there is little or no relationship between these variables and friends' maternal education for male students. On the other hand, the smoking/drinking index is associated with maternal college for male students, and physical health is associated with maternal drop-out for the female sample while these health behavior oriented variables have little or no relationship maternal college graduate for the female sample.

Our results support that girls are more influenced by high quality peers than boys on self evaluation and social comfort, as suggested in relevant previous literature (Brown 1982, Griffin et al., 1999), but less likely to be influenced in terms of exhibiting problematic behaviors. We also run correlation analysis and confirm that low self evaluation, passive attitude and behavior at school and poor mental health are negatively associated with GPA in our sample, even after removing school by cohort and school by student type fixed effect.

As a further test of the relevance of these potential mechanism variables, we re-estimate our two-stage IV model for girls allowing the effect of predicted friendships to vary across the three maternal education subgroups: maternal college educated, maternal high school graduate and maternal high school drop-out. The resulting estimates are shown in the first column of Table 10 and imply that most of the effect of maternal college friends on GPA is concentrated in the maternal college educated subsample with an effect of 0.279 approximately 42 percent larger than the estimate for the full sample. We observe a statistically insignificant positive effect of

[26] Self-evaluation covers rating to questions including whether the interviewees think themselves physically fit, are proud of themselves, like themselves, think they are doing things right, and try to study well.
[27] Environment evaluation shows the extent that students feel close, safe, fair and accepted at school.

0.105 and no effect for the maternal high school drop-out subsample. We then re-estimate the models for the mechanism variables finding that the positive effects of maternal college on most of these variables (social comfort, mental health and misbehave) is also concentrated among the maternal college sample, and to a much lesser extent in the maternal high school graduate sample, with fewer effects in the maternal drop-out sample.

The terms self evaluation and subjective feeling on social environment here fall loosely in the concept of general self-esteem or self-concept, which are important indicators for troublesome behavior and depression (Rosenberg et al. 1989, Markowitz 2001).[28] Numerous studies in education have found that academic achievement and self-esteem are positively correlated (see Bankston & Zhou, 2002; Ross & Broh, 2000; Schmidt & Padilla, 2003; Wong & Watkins, 2001). Purky (1970) argued that there is continuous interaction between self-esteem and academic achievement. Byrne (1984) reviewed the empirical findings in this literature, both cross-sectional and longitudinal designs, and also confirmed the existence of the relationship. Our analysis is novel because we have plausibly exogenous variation in friendship composition that can separate correlational and causal effects. However, the causal link between self-esteem and school achievement is still under debate. Some investigators argue that high school achievement and self control enhance self esteem, not vice versa, and our analysis cannot shed light on this debate because, as with the mechanism analyses in earlier cohort studies (Bifulco et al. 2011, Lavy and Schlosser 2011), we have only shown that friendship composition has a causal influence on both the variable of interest and on the potential mechanism and not whether one of these effects operates through the other.

Simulated Effects on Friendship Patterns and GPA

Next we conduct some simple calculations and simulations in order to assess the impact of changing school composition of maternal education on both the friendship composition and on the educational outcomes of girls in our sample. We begin with a simple calculation of the effect of our key significant findings on the effect of educational composition on friendship formation. Specifically, in Table 4, we show that an increase in the share of maternal college educated students increases the likelihood of a match between two students who both have a college educated mother as well as the likelihood of a match between two students where one has a college educated mother and the other has a mother who is a high school graduate. In our sample, we increase the share of maternal college educated students in every cohort by 10 percentage points and then examine the direct effect of this increase on the predicted number of college educated for students overall and by level of maternal education.

The predicted number of maternal college educated friends changes with number or share of maternal college students for two reasons: 1. There are simply more potential friendship matches available with students whose mothers are college graduates, and 2. The probability of matches or links increases both between two maternal college students and between a maternal college and a maternal high school graduate student based on the statistically significant estimates on percent maternal college in Table 4.

[28] When discussed in Psychology, the concept of self-esteem often needs to be clarified—either a general term on overall feeling about self or on a specific aspect, such as academic related, physical appearance and social popularity, etc.

Assuming that cohort size is held constant, the number of maternal college educated students after the change (N_{2c}) is simply

$$N_{2c} = N_{1c} + 0.1N \tag{19}$$

where N_{1c} is the initial number of maternal college educated students, and N is the total number of students in the cohort. The resulting change in number of maternal college friendship links or matches for a maternal college educated student (D_c) is then

$$D_c = (N_{1c} + 0.1N - 1)\left(P_{cc} + \beta_{cc}\frac{0.1}{N}\right) - (N_{1c} - 1)P_{cc} \tag{20}$$

where P_{cc} is the probability of a link and β_{cc} is the estimated coefficient on the share maternal college educated for college-college links. This expression can be rewritten to illustrate the separate effects of changes in the probability of a link and changes in the number of potential links

$$D_c = (N_{1c} + 0.1N - 1)\left(\beta_{cc}\frac{0.1}{N}\right) + (0.1NP_{cc}) \tag{21}$$

For maternal high school graduate students the change in maternal college friends (D_h) is

$$D_h = (N_{1c} + 0.1N)\left(\beta_{ch}\frac{0.1}{N}\right) + (0.1NP_{ch}) \tag{22}$$

where P_{ch} is the probability of a link and β_{ch} is the estimated coefficient on the share maternal college educated for college-high school links. Finally, for maternal high school drop-out students we set the parameter estimate on share maternal college to zero due to the small and statistically insignificant estimate and the predicted change is

$$D_d = 0.1NP_{cd} \tag{23}$$

where P_{cd} is the probability of a link.

In order to calculate these expressions for the sample, we use the within school sample average frequencies of link formation between potential links for P_{cc}, P_{ch} and P_{cd}. We set P_{cc}, P_{ch} and P_{cd} to the empirical frequencies observed in each school so that our calculations capture the fact that schools differ in the likelihood of link formation due to, for example, across school differences in the racial and ethnic composition of each maternal education subgroup. Further, the use of the empirical frequencies is consistent with holding cohort size constant because one would expect link frequencies to fall on average as cohort size increases. These results are shown in Column 1 of Table 11 where the rows present the results for the overall sample, the maternal college subsample, the maternal high school subsample and the maternal drop-out subsample. A 10 percentage point increase in the share of maternal college students increases the sample average fraction of students who have a mother with a college degree by 34 percent over an original base fraction of 0.295.[29] The average number of predicted maternal college friends

[29] The increase is 41% if the fraction of maternal college students is calculated based on all four maternal education categories: college, high school, drop-out and missing, rather than omitting missing from the calculation.

increases by 0.234 from a base of 0.310 or by 75 percent. The maternal college subsample has an increase of 0.240 over a base of 0.532 or 45%, and the maternal high school graduate sample has an increase of 0.313 over a base of 0.293 or 107%. The maternal high school drop-out sample has the smallest absolute increase of 0.136, but the largest percent increase of 111% over a base of 0.123.

Notably, the percentage increases in maternal college friends are substantially smaller for maternal college educated students than for maternal high school graduate or drop-out students. While this seems surprising given the strong effects of percent of students with maternal college graduates on the likelihood of link formation between two students with maternal college graduates, the increase in the likelihood of college-college and college-high school links explains only a moderate fraction of the increased average number of friends, 0.049 for the maternal college subsample and 0.042 for the maternal high school graduate subsample.[30] The primary driver of the increase in the number of maternal college friends for all groups is the increasing number of friendship opportunities. This effect is smallest for maternal college students because the percent increase in maternal college students is smallest in the cohorts that have the largest share of maternal college students. In the maternal college subgroup, observations are more likely to come from schools with a larger share of maternal college students than average and so the smallest percent increase in maternal college friendship opportunities. Further, given the strong negative correlation between the presence of maternal college students and maternal high school drop-out students, the largest percentage increases in maternal college friendship opportunities occur for the maternal high school drop-out subsample in the schools with the largest maternal high school drop-out population.

Next, we conduct a series of simulation analyses where we use the entire estimated friendship formation model to predict changes in friendship composition as we increase the share of maternal college students by 10 percentage points in each cohort while holding the size of the cohort fixed, e.g. by dropping maternal high school graduate, maternal high school drop-out and maternal education missing students with probabilities based on their relative shares within cohort as maternal college students are added.[31] As noted above, this change increases the percent maternal college in the sample by 34 percent. The simulations follow two distinct scenarios. The first scenario assigns the race and ethnicity of the added maternal college students based on the racial and ethnic composition of the original population of maternal college students in each cohort. As a result, cohort percent African-American and percent Hispanic decrease on average, especially percent Hispanic, because these groups are more heavily represented among maternal high school graduate and maternal high school drop-out. These simulations are conducted 30 times and then the averages of the simulations are presented.[32]

The simulation results for this scenario are presented in column 2 of Table 11. The increase in share maternal college graduate leads to an 0.1 percentage point or 0.5% decline in percent African-American and an 0.8 percentage point or 4.8% decline in percent Hispanic. The

[30] Given the base numbers of friends above, the percent increase for maternal college students is 9 percent and the percent increase for maternal high school graduate students is 14 percent comparable to, but somewhat larger than, the back of the envelop calculations presented earlier of 7 and 10 percent. The large estimates arise in part because the effect of the increased probability of link formation is calculated for a larger number of maternal college students as shown in the first terms of equations (21) and (22).

[31] The increase in percent maternal college education is sufficiently large to change the implied weight on each cohort for the maternal education subsample means. Reweighting so that the mean represents the effects given the original maternal educational composition of the sample, however, has very little impact on our simulation results.

[32] Nearly identical results arise when the simulations are conducted 100 times.

average number of predicted maternal college friends increases by 0.175 from a base of 0.310 or by 56 percent. For the maternal college subsample, the increase is 0.109 over a base of 0.532 or 20%. The maternal high school graduate sample has an increase of 0.180 over a base of 0.293 or 61%, and maternal high school drop-out sample has an increase of 0.128 over a base of 0.123 or 104%.

The simulated changes are substantially smaller for the maternal college and maternal high school subsample than in the calculations while the simulated change for maternal drop-out declines only slightly. As discussed earlier, decreases in percent African-American and percent Hispanic reduce the likelihood of homophilous friendship links and increase the likelihood of heterogenous friendship links. Both the maternal college and high school subsamples are predominantly white and so a decline in the likelihood of friendship links between whites will reduce the number of links with maternal college students for both subsamples. This effect is primarily driven by Hispanics given the lower levels of maternal education in the Hispanic subsample. A second reason behind these declines is the effect of the large, but statistically insignificant, coefficients on maternal high school drop-out for all links involving at least one maternal college student. This effect is especially large for maternal college students because in cohorts where there are a large number of maternal college graduate students a disproportionately large numbers of maternal high school graduate and high school drop-out students must be dropped. Note that for maternal drop-outs these effects work in opposite directions because the increase in the likelihood of racially heterogenous friendship links should tend to increase the number of maternal college friends for the drop-out subsample consistent with the smaller decline for this subsample.

The second simulation scenario assigns each maternal college student who is added to a cohort the race and ethnicity of the non-maternal college student who is dropped from the school in order to keep cohort racial composition constant. The average number of predicted maternal college friends increases by 0.163 from a base of 0.310 or by 53 percent. For the maternal college subsample, the increase is 0.106 over a base of 0.532 or 20%. The maternal high school graduate sample has an increase of 0.161 over a base of 0.293 or 55%, and maternal high school drop-out sample has an increase of 0.118 over a base of 0.123 or 96%.

In this case, the racial and ethnic composition of the cohort is unchanged, but the racial and ethnic composition of the maternal college subsample is changed. Specifically, we increase the diversity of the maternal college cohort by adding, on average, individuals who are drawn from a less white subsample of students. As with the previous scenario, this effect leads to reductions in the likelihood of college-college links and college-high school links relative to the calculations, and so decreases the increase in the number of maternal college links for each group. On the other hand, for the maternal drop-out sample, the number of maternal college friends declines because the effect of lower share minority in the previous scenario has been eliminated.

In terms of estimating the impact on GPA, the first stage effect of predicted number of friends on actual friends is 0.844, and then the effect on GPA is 0.196 from the instrumental variable analysis. Therefore, our calculations suggest that the direct effect associated with a 10 percentage point increase in the share of maternal college students in each cohort increases girls GPA by 0.039.[33] For the simulation results, which allow for racial and other compositional

[33] This result arises from multiplying the predicted change in number of maternal college educated friendships times the first stage instrumental variables coefficient in order to obtain the correct scale and then multiplying this product

effects on friendship formation, the predicted increases for the full sample are smaller at between 0.029 and 0.027. Turing to the maternal education subsamples, we draw on the estimated effects in Table 10 where we observed large positive and significant results for the maternal college subsample and insignificant, but appreciable, effects for the maternal high school subsample. For the calculated increases in number of maternal college friends, the increases in GPA are 0.040 for maternal college students and 0.052 for maternal high school students. Similarly, the predicted effects from the simulations are smaller at 0.018 (same for both scenarios) for maternal college students and between 0.030 and 0.027 for maternal high school students.

Conclusions

This paper presents new evidence of the determinants of friendship links and the effects of the characteristics of friends on own school achievement. We use a novel strategy that leverages across-cohort, within school variation in the "supply of friend 'types'" for both sets of results. We first show that small variations in the supply of friends increase homophily and reduce heterophily in friendship formation patterns in high school. This is consistent with both the biological evidence that individuals prefer to have friends like themselves as well as the large body of empirical work that shows strong correlations in the characteristics of friends (i.e. homophily). However, we are the first to examine these effects within a quasi-experimental research design[34] and to provide evidence of how the pattern of homophily increases as the population of minority groups increase. These results have strong implications for policies that attempt to "rewire" social networks by increasing the opportunities for choosing friends who are different. Our results suggest that increasing opportunities may not be enough to foster heterophilious friendships, which is also a likely explanation for the results from Carrell et al. (2011), where randomizing individuals into squadrons in the military to foster heterophily actually reduced the performance of the heterophilious group.

We then use our predictions of friendship formation to leverage a second research question—whether having friends with highly educated mothers is causally related to academic achievement or whether the correlation in GPA between friends is a result of endogenous friendship selection. We find both cases—for female high school students, our results suggest that increases in friend maternal education status leads to large GPA increases, which are concentrated in coursework in science and English. We also find that the OLS estimates for males are driven by endogeneity, and once corrected are small and no longer statistically significant.

In order to examine the mechanisms linking the maternal education of friends to own academic achievement, we show evidence that, for females but not males, friend maternal education is also linked to reductions in increases in feelings of self worth and favorable opinions of the school environment. Having more friends with college educated parents appears to lead to both higher levels of self-esteem and higher grades among girls while friendship composition does not appear to affect other significant intermediate outcomes like disciplinary problems or health outcomes. We also show that both our findings for GPA and our findings for the mechanism variables are concentrated among the subsample of students whose mothers have a college degree.

---

by the coefficient on GPA from the second stage model or 0.234*0.844*0.196=0.034. The subsample calculations follow the same form except use the change in number of maternal college educated friends for each subsample.

[34] The most similar work examines dating patterns rather than high school friendship formation (Fisman et al. 2008).

Finally, we conduct a series of calculations and simulations in order to illustrate the effect of small to moderate increases in share of students with maternal education at the college level. The key findings from these simulations are 1. that the increased opportunities for additional friendships with maternal college educated students dominates the direct effect associated with changes in the probability of friendship formation as the share of maternal college students increases, 2. the opportunity effects are largest among students with less educated mothers where the percent increase in share maternal college are largest and 3. that these increases in the number of friendships for students with maternal college graduates are reduced substantially for the maternal college and maternal high school subsamples when we use the entire estimated friendship formation model so that friendship patterns are allowed to respond to general changes in composition of students in cohorts including for example changes in the racial and ethnic composition.

References

Ballester, C., Calvó-Armengol, A. and Y. Zenou (2006), "Who's who in networks. Wanted: the key player", *Econometrica*, 74, 1403-1417.

Bankston, C. L. and M. Zhou, (2002). "Being Well vs. Doing Well: Self-esteem and School Performance among Immigrant and Non-immigrant Racial and Ethnic Groups." *International Migration Review*, 36, 389-415.

Bayer, P., Ross, S. and G. Topa. 2008. "Place of work and place of residence: Informal hiring networks and labor market outcomes." *Journal of Political Economy*, 116 (6): 1150-1196

Bifulco, R., Fletcher, J. and S. Ross. 2011. "The effect of classmate characteristics on individual outcomes: evidence from the Add Health." *American Economic Journal: Economic Policy* 3(1):25–53

Billings, S. B., Deming, D. J. and Jonah E. Rockoff. (2012). "School Segregation, Educational Attainment and Crime: Evidence from the end of busing in Charlotte-Mecklenburg." NBER Working Paper 18487

Brown, B.B. (1982). "The Extent and Effects of Peer Pressure among High School Students: a Retrospective Analysis." *Journal of Youth and Adolescence*. 11(2), 121-133.

Byrne, Barbara M. (1984). "The General/Academic Self-concept Nomological Network: a Review of Construct Validation Research." *Review of Educational Research.* 54(3): 427-456.

Calvó-Armengol, A., Patacchini, E. and Y. Zenou (2009), "Peer effects and social networks in education", *Review of Economic Studies*, 76, 1239-1267.

Carrell, Scott, Bruce Sacerdote, and James West (in press). "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation," Econometrica

Currarini, Sergio, Matthew Jackson, and Paolo Pin. (2009). "An Economic Model of Friendship: Homophily, Minorities, and Segregation." Econometrica, 77(4): 1003-1045

Durlauf, S. 2004. "Neighborhood effects." In *The Handbook of Regional and Urban Economics, Volume 4: Cities and Geography*, edited by V. Henderson and J.F. Thisse. Elsevier Science/North Holland.

Fisman, Raymond, Sheena Iyengar, Emir Kamenica, and Itamar Simonson. (2008). Racial Preferences in Dating. Review of Economic Studies 75(1): 117-132

Fletcher, Jason M. and Stephen L. Ross (2012). Estimating the Effects of Friendship Networks on Health Behaviors of Adolescent. NBER Working Paper 18253

Griffin, K. W., Botvin, G. J., Doyle, M. M., Diaz, T., & Epstein, J. A. (1999). "A six-year follow-up study of determinants of heavy cigarette smoking among high school seniors." *Journal of Behavioral Medicine*, 22, 271-284.

Guryan, J, D Kroft and N J. Notowidigdo. 2009. **"**Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics*, 1, 34-68.

Hoxby, Caroline. 2000. "Peer Effects in the Classroom: Learning from Gender and Race Variation." National Bureau of Economic Research Working Paper 7867.

Jackson, M.O. Social and Economic Networks. Princeton, NJ: Princeton University Press, 2008.

Lavy, VD and A Schlosser. 2011. "Mechanisms and impacts of gender peer effects at school." *American Economic Journal: Applied Economics.* 3(2): 1–33

Lazarfeld, P. F., and R. K. Merton (1954): "Friendship as a Social Process: A Substantive and Methodological Analysis," in *Freedom and Control in Modern Society*, ed. by M. Berger. New York: Van Nostrand, 18–66.

Manski, C. 1993. "Identification of endogenous social effects: The reflection problem." *Review of Economic Studies*. 60

Manski, C. 1995. Identification Problems in the Social Sciences. Harvard University Press: Cambridge, MA

Manski, C. 2000. "Economic analysis of social interactions." *Journal of Economic Perspectives*. 14:3

Markowitz, F. E. (2001). "Modeling processes in recovery from mental illness: Relationships between symptoms, life satisfaction, and self-concept." *Journal of Health and Social Behavior*, 42, 64–79.

Mayer, Adalbert and Steven Puller (2008). The old boy (and girl) network: Social network formation on university campuses. Journal of Public Economics 92 (1-2), 329-347

Montgomery, James. (1991). Social Networks and Labor-Market Outcomes: Towards an Economic Analysis. American Economic Review, 81(5): 1408-1418

Moody, J. (2001): "Race, School Integration, and Friendship Segregation in America," *American Journal of Sociology*, 107, 679–716.

Pomerantz, E. M., Altermatt, E. R., & Saxon, J. L. (2002). "Making the Grade but Feeling Distressed: Gender Differences in Academic Performance and Internal Distress." *Journal of Educational Psychology*, 94, 396-404.

Purky, W. (1970). *Self concept and school achievement.* New Jersey: Prentice-Hall.

Rosenberg, M., Schooler, C., and Schoenbach, C. (1989). "Self esteem and adolescent problems: Modeling reciprocal effects." *American Sociological Review*. 54, 1004–1018.

Ross, C. E., & Broh, B. A. (2000). "The Roles of Self-esteem and the Sense of Personal Control in the Academic Achievement process." *Sociology of Education,* 73, 270-284.

Schmidt, J. A., & Padilla, B. (2003). "Self-esteem and Family Challenge: An Investigation of Their Effects on Achievement." *Journal of Youth and Adolescence,* 32, 37-46.

Weinberg, Bruce. (2007). "Social Interactions with Endogenous Associations." NBER Working Paper 13038

Wong, M. S. W, & Watkins, D. (2001). "Self-esteem and Ability Grouping: A Hong Kong Investigation of the Big Fish Little Pond Effect." *Educational Psychology: An International Journal of Experimental Educational Psychology*, 21, 79-87.

Results Tables

Table 1. Number and distribution of mutual friends by maternal education

		Friend's Maternal Education			
	No. of Friends	High School Dropout	High School Graduate	College Graduate	Missing
Own Maternal Education					
Female					
High School Dropout	0.878	23.4%	45.9%	14.1%	16.6%
High School Graduate	1.149	9.3%	53.7%	25.8%	11.1%
College Graduate	1.282	4.9%	43.8%	42.1%	9.2%
Missing	0.782	13.6%	45.5%	22.1%	18.8%
Male					
High School Dropout	0.513	17.3%	45.8%	16.2%	20.7%
High School Graduate	0.717	7.0%	50.1%	28.9%	14.1%
College Graduate	0.848	3.3%	39.6%	45.9%	11.2%
Missing	0.461	9.5%	43.2%	24.7%	22.6%
	N				
Maternal Education Distribution	84654	10.56%	44.48%	25.44%	19.51%

Note: The “No. of friends” column presents mean of the number of mutual friends occurred to a student by gender in each category of maternal education. A mutual friend tie is defined as a two-way nomination of friendship. The last row includes the total sample size and distribution by maternal education. In the rest of the tables, we will label high school dropout as “HS dropout”, high school graduate as “HS grad”, and college graduate as “College Grad” when space is limited.

Table 2. Number and distribution of mutual friends by racial groups

			Friend's Race			
		No. of Friends	White	Black	Hispanic	Other
Own Race						
Female						
	White	1.266	87.8%	1.6%	5.3%	5.4%
	Black	0.883	7.0%	80.7%	7.2%	5.0%
	Hispanic	0.790	28.6%	8.7%	55.1%	7.6%
	Other	0.949	41.3%	8.9%	10.6%	39.2%
Male						
	White	0.822	86.0%	1.7%	5.6%	6.7%
	Black	0.445	11.0%	73.7%	9.4%	5.8%
	Hispanic	0.455	32.7%	8.4%	49.5%	9.5%
	Other	0.617	43.3%	5.7%	10.9%	40.2%
		N				
Race Distribution		84654	55.72%	16.83%	16.89%	10.55%

Note: The “No. of friends” column presents mean of the number of mutual friends occurred to a student by gender in each category of racial group. A mutual friend tie is defined as a two-way nomination of friendship. The last row includes the total sample size and distribution by racial groups.

Table 3. Balancing test for cohort composition sorting with student demographic characteristics

Variables	%Black (1)	%Hispanic (2)	%Mom College Grad (3)	%Mom HS Dropout (4)
Age	-0.00040	-0.00023	0.00003	0.00061*
	(0.00041)	(0.00039)	(0.00040)	(0.00028)
No. of People	0.00006	-0.00010	0.00007	0.00003
in Household	(0.00013)	(0.00015)	(0.00025)	(0.00017)
No. of School Kids	0.00010	0.00026	0.00010	0.00017
in Household	(0.00022)	(0.00017)	(0.00026)	(0.00019)
Live with Both	-0.00013	0.00001	0.00058	-0.00049
Parents	(0.00059)	(0.00045)	(0.00065)	(0.00044)
Live with Biological	0.00052	0.00019	0.00109	-0.00032
Parents	(0.00168)	(0.00141)	(0.00261)	(0.00173)
Mother's Edu in	-0.00001	-0.00010	0.00032	-0.00010
Single Year	(0.00014)	(0.00008)	(0.00020)	(0.00011)
Mother Born in US	0.00213	-0.00024	-0.00012	0.00018
	(0.00149)	(0.00060)	(0.00081)	(0.00067)
Born in US	-0.00124	-0.00016	-0.00148	0.00040
	(0.00101)	(0.00092)	(0.00103)	(0.00081)
Adopted	-0.00070	0.00028	0.00153	-0.00152
	(0.00136)	(0.00144)	(0.00171)	(0.00145)
Health Condition	0.00043	-0.00014	-0.00048	-0.00117
at Birth	(0.00110)	(0.00129)	(0.00149)	(0.00108)
Observations	56,774	56,774	56,769	56,769
R-squared	0.974	0.972	0.908	0.877
Ftest	1.001	0.901	0.748	0.991
Fpvalue	0.446	0.534	0.678	0.454

Note: Each column displays a separate regression of a cohort composition variable on ten predetermined demographics variables. The cohort composition variables for a student includes the percentage of black (not Hispanic), Hispanic, mother graduated from four year college and mother dropout from high school, omitting the student's contribution. All regressions control for school-gender fixed effect, grade dummies, and Guryan type control for school level composition omitting the student him/herself. Standard errors are clustered at the school level. Observations with missing maternal education data are assigned the median value of the cohort variable of all other students in the school-grade-gender group. **p<0.01 and *p<0.05.

Table 4. Friendship Pattern Estimation from Matching Model
Estimates for Maternal Education

	Female		Male	
	% College Grad	% HS Dropout	% College Grad	% HS Dropout
Maternal Education Pair Type				
College Grad-College Grad	1.242**	2.202	0.228	3.066
	(0.323)	(2.872)	(0.191)	(1.796)
College Grad-HSGrad	1.291**	1.688	0.478	0.636
	(0.231)	(1.438)	(0.352)	(0.708)
College Grad-HS Dropout	0.068	0.683	1.110	0.013
	(0.968)	(0.869)	(0.709)	(0.011)
HSGrad-HSGrad	0.626	0.751	0.299	0.212
	(0.391)	(0.493)	(0.297)	(0.799)
HSGrad-HS Dropout	1.183	0.674	0.206	0.036
	(0.677)	(0.604)	(0.608)	(0.274)
HS Dropout-HS Dropout	-0.880	-0.754	1.367	-0.656
	(1.734)	(0.699)	(1.008)	(0.945)

Note: Each column and row displays coefficient from separate regression of an indicator of whether or not two students are mutual friend on a series of interaction terms between the variable of the column and the variable of the row. A student pair sample of 6 million observations is estimated, which is composed of all unique pair of students within a school by grade by gender cell. The dependent variable is binary indicator of whether the two students in a pair both nominate each other as their friend. The independent variables are interactions between binary indictors for the type of a student pair and cohort composition variables. Maternal education type (defined by the match of the two parties' maternal education, as in each row, e.g. college grad-college grad) is interacted with percentage of college graduate mothers and the percentage of high school dropout mothers weighted by cohort size by gender. Racial type is interacted with percentage of black students and percentage of Hispanic students. This table presents coefficients from maternal education interactions. Coefficients of types with missing maternal education from one or both parties are omitted in this table because of ambiguous implication. All regressions control for school-gender-cross pair type fixed effect. The cross pair type combines maternal education and race, e.g. white-high school grad-black-college grad. Standard errors are clustered at the school level. **p<0.01 and *p<0.05.

Table 5. Friendship Pattern Estimation from Matching Model
Estimates for Race/Ethnicity

	Female		Male	
Racial Pair Type	% Black	% Hispanic	% Black	% Hispanic
White-White	1.795**	1.604	2.370**	1.240
	(0.433)	(0.848)	(0.773)	(0.661)
White-Black	-0.012	-0.161	-0.044	0.365
	(0.158)	(0.406)	(0.216)	(0.908)
White-Hispanic	-0.500	-0.863*	0.196	-0.024
	(1.150)	(0.401)	(0.660)	(0.285)
White-Other	-1.353*	-0.563	1.176	0.581
	(0.677)	(0.665)	(0.827)	(0.629)
Black-Black	0.578	-1.603	0.322	2.757**
	(0.316)	(1.501)	(0.212)	(0.998)
Black-Hispanic	0.230	-0.251	0.099	-0.089
	(0.449)	(0.410)	(0.255)	(0.279)
Black-Other	0.065	-2.184*	-0.122	0.165
	(0.121)	(1.086)	(0.106)	(0.630)
Hispanic-Hispanic	-2.066	0.822	1.767	0.128
	(1.850)	(0.598)	(1.245)	(0.358)
Hispanic-Other	-2.600*	0.555	0.977	0.004
	(1.206)	(0.502)	(0.894)	(0.300)
Other-Other	9.456**	1.673	-0.994	-0.795
	(3.159)	(1.976)	(1.303)	(1.538)

Note: Each column and row displays coefficient from separate regression of an indicator of whether or not two students are mutual friend on a series of interaction terms between the variable of the column and the variable of the row. This table presents coefficients of racial interactions from the same regression described in the notes of table 4. See notes of table 4. All regressions control for school-gender-cross pair type fixed effect. Standard errors are clustered at the school level. **p<0.01 and *p<0.05.

Table 6. Distribution of Maternal Education and GPA by Race

	White	Black	Hispanic	Other
Mother High School Dropout	8.4%	10.6%	34.5%	13.2%
Mother High School Graduate	57.9%	58.8%	47.2%	44.9%
Mother College Graduate	33.7%	30.6%	18.2%	41.8%
Average GPA for females	3.018	2.683	2.661	3.077
Average GPA for males	2.868	2.487	2.522	2.890

Note: The first three rows are the percentage of each maternal education category by race. Observations with missing maternal education are not counted.

Table 7. Effect of Friends Pattern on Student's GPA

	Female			Male		
	OLS	IV1	IV2	OLS	IV1	IV2
No. of Friends with	0.152**	0.196*		0.154**	-0.017	
Mom College Grad	(0.010)	(0.087)		(0.013)	(0.219)	
No. of Friends with	-0.010		0.007	-0.030		0.063
Mom High School Dropout	(0.014)		(0.140)	(0.024)		(0.099)
Guryan Control		0.016			-0.059	
Mom College Grad		(0.047)			(0.145)	
Guryan Control			-0.043			-0.088
Mom High School Dropout			(0.097)			(0.085)
R-squared	0.245	0.042	0.027	0.231	0.022	0.024
N	36903	36660	36660	35373	35116	35116
Type*School FE	Yes	Yes	Yes	Yes	Yes	Yes
Grade*School FE	Yes	Yes	Yes	Yes	Yes	Yes
Weak IV F-stat		71.980	62.732		29.905	246.671
Anderson-Rubin Wald test F		5.06	0.00		0.01	0.41
P-val		0.026	0.961		0.938	0.525
10% Likelihood of Rej		16.38	16.38		16.38	16.38

Note: Each column displays a separate regression of GPA on number of mutual friends with college graduate mothers and/or with high school dropout mothers. In IV regression, numbers of mutual friends are instrumented with corresponding predicted number of friends. All regressions control for school-gender-cross pair type fixed effect and school-grade fixed effect. IV regressions include Guryan type control for school level friendship pattern in both first and second stage. Standard errors are clustered at the school level. **p<0.01 and *p<0.05.

Table 8. Effect of Friends Pattern on Student's GPA by Subject

	Female				Male			
	Math	English	Science	History	Math	English	Science	History
No. of Friends with	-0.021	0.276*	0.354**	0.133	-0.213	-0.012	-0.033	0.186
Mom College Grad	(0.135)	(0.121)	(0.124)	(0.101)	(0.285)	(0.290)	(0.295)	(0.245)
R-squared	0.024	0.034	0.028	0.051	-0.010	0.025	0.026	0.045
Weak IV F	71.345	66.904	59.837	63.131	33.337	26.143	21.221	35.715
No. of Friends with	0.098	0.175	-0.009	-0.115	-0.112	0.061	0.086	0.030
Mom HS Dropout	(0.163)	(0.187)	(0.183)	(0.187)	(0.151)	(0.128)	(0.146)	(0.153)
R-squared	0.025	0.024	0.036	0.038	0.021	0.026	0.029	0.037
Weak IV F	62.827	79.778	77.288	67.737	214.51	206.89	192.69	170.93
N	34499	35490	32381	32096	33463	34050	31291	30982

Note: Each column and row displays a coefficient from a separate IV regression. All regressions control for school-gender-cross pair type fixed effect and school-grade fixed effect. Guryan type control for school level friendship pattern is included in both first and second stage. Standard errors are clustered at the school level. **p<0.01 and **p<0.05.

Table 9. Mechanism Analysis

	Evaluate	Environ	Mental	Trouble	Misbehave	Addict	Health
Female							
No. of Friends with	0.261*	0.189#	0.181#	-0.014	-0.191#	-0.067	0.036
Mom College Grad	(0.115)	(0.102)	(0.099)	(0.090)	(0.105)	(0.087)	(0.141)
No. of Friends with	0.004	0.234	-0.003	-0.032	0.082	0.147	-0.283#
Mom HS Dropout	(0.168)	(0.143)	(0.163)	(0.142)	(0.159)	(0.150)	(0.168)
N	36853	36208	37278	39943	37565	39029	37406
R-squared	0.012	0.039	0.017	0.018	0.007	0.035	0.015
Male							
No. of Friends with	0.042	0.075	0.228	-0.126	-0.288	-0.385#	0.110
Mom College Grad	(0.190)	(0.197)	(0.199)	(0.220)	(0.236)	(0.228)	(0.179)
No. of Friends with	0.009	0.145	-0.010	0.014	-0.049	-0.151	-0.152
Mom HS Dropout	(0.142)	(0.155)	(0.132)	(0.143)	(0.147)	(0.125)	(0.144)
N	34892	34257	35268	38336	35836	37447	35448
R-squared	0.016	0.024	0.014	0.017	0.005	0.008	0.015

Note: Each column and row displays a coefficient from a separate IV regression. Dependent variables are constructed by factor analysis of students' report on own mental status, behavior, school and family environment (see Appendix Table 9A for reference). All regressions control for school-gender-cross pair type fixed effect and school-grade fixed effect. Guryan type control for school level friendship pattern is included in both first and second stage. Standard errors are clustered at the school level. **p<0.01, *p<0.05, #p<0.1.

Table 10. Female Sample IV Results of Maternal College Friends Interacted with Own Maternal Education

	GPA	Evaluate	Environ	Mental	Trouble	Misbehave	Addict	Health
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
College Graduate	0.279**	0.283#	0.250*	0.261#	-0.072	-0.307#	-0.105	0.194
	(0.100)	(0.147)	(0.124)	(0.147)	(0.105)	(0.167)	(0.124)	(0.164)
High School Graduate	0.105	0.217#	0.098	0.118	0.087	-0.107	-0.118	-0.224
	(0.113)	(0.120)	(0.134)	(0.122)	(0.133)	(0.131)	(0.131)	(0.154)
High School Dropout	-0.022	0.387#	-0.013	-0.097	-0.195	-0.014	-0.007	0.025
	(0.217)	(0.212)	(0.318)	(0.224)	(0.245)	(0.257)	(0.255)	(0.292)
N	36660	36853	36208	37278	39943	37565	39029	37406
R-sq	0.032	0.010	0.031	0.011	0.011	0.002	0.031	0.007

Note: Each column displays the coefficients from a separate IV regression. Interaction terms of actual number of maternal college friends and four dummies for own maternal education respectively are instrumented with predicted number of maternal college friends interacted with four dummies for own maternal education. The coefficients for the interaction term with missing maternal education are not presented. Dependent variables except GPA are constructed by factor analysis of students' report on own mental status, behavior, school and family environment (see Appendix Table 9A for reference). All regressions control for school-gender-student type fixed effect and school-grade fixed effect. Guryan type controls for school level friendship pattern (also interacted with dummies for own maternal education) are included in both first and second stage. Standard errors are clustered at the school level. **p<0.01, *p<0.05, #p<0.1.

Table 11. Simulation Results for Female

Own Maternal Education	Baseline	Increase in Number of College Friends		
		Calculation	Simulation Scenario 1	Simulation Scenario 2
All	0.310	0.234	0.175	0.163
College Graduate	0.532	0.240	0.109	0.106
High School Graduate	0.293	0.313	0.180	0.161
High School Dropout	0.123	0.136	0.128	0.118

Note: Baseline is average of actual number of friends with maternal college education for the whole female sample and for female subsamples by own maternal education respectively. The other columns represent the increase in average number of maternal college friends relative to the baseline.

Appendix Tables

Table 1A. Average Number of Mutual Friends, Deviation from School Share

		Friend's Maternal Education			
	No. of Friends	HS Dropout	HS Grad	College Grad	Missing
Own Maternal Education					
Female					
High School Dropout	0.878	0.072	0.020	-0.055	-0.029
High School Graduate	1.149	-0.027	0.035	0.017	-0.059
College Graduate	1.282	-0.060	0.010	0.091	-0.066
Missing	0.782	0.012	0.023	-0.005	-0.020
Male					
High School Dropout	0.513	0.052	0.033	-0.101	-0.035
High School Graduate	0.717	-0.013	0.055	0.027	-0.085
College Graduate	0.848	-0.047	0.036	0.094	-0.102
Missing	0.461	0.000	0.021	0.007	-0.023
Maternal Edu Distribution	1.000	10.56%	44.48%	25.44%	19.51%

Note: We calculate the friendship frequencies in Tables 1 and 2 for every school and subtract the school fraction of the friendship type (columns) in order to get a deviation from expected based on frequency. A weighted average of this across all schools with weights based on number of students by type (rows) will deliver the empirical level of within school homophily.

Table 2A. Average Number of Mutual Friends, Deviation from School Share

		Friend's Race			
	No. of Friends	White	Black	Hispanic	Other
Own Race					
Female					
White	1.266	0.136	-0.342	-0.068	-0.070
Black	0.883	-0.327	0.329	-0.255	-0.055
Hispanic	0.790	-0.047	-0.074	0.171	-0.049
Other	0.949	-0.004	-0.094	-0.082	0.156
Male					
White	0.822	0.137	-0.261	-0.076	-0.078
Black	0.445	-0.292	0.295	-0.114	-0.021
Hispanic	0.455	-0.014	-0.014	0.132	-0.048
Other	0.617	0.019	-0.080	-0.054	0.151
Race Distribution	1.000	55.72%	16.83%	16.89%	10.55%

Note: We calculate the friendship frequencies in Tables 1 and 2 for every school and subtract the school fraction of the friendship type (columns) in order to get a deviation from expected based on frequency. A weighted average of this across all schools with weights based on number of students by type (rows) will deliver the empirical level of within school homophily.

Table 3A. Predicted vs. Actual Friendship Patterns

		Actual Number of Friends		Predicted Number of Friend	
		Mean	Std. Dev.	Mean	Std. Dev.
Total	84654	0.882	1.144	0.868	0.530
Mom High-School Dropout	84654	0.077	0.298	0.076	0.109
Mom College Graduate	84654	0.269	0.611	0.265	0.266
Mom's Education Missing	84654	0.116	0.355	0.114	0.116
Black Friends	84654	0.115	0.449	0.113	0.257
Hispanic Friends	84654	0.105	0.389	0.103	0.174
Other Race Friends	84654	0.081	0.340	0.079	0.159

Table 4A. Frequency of Actual Ties among All Potential Ties

	Female	Male		Female	Male
Racial Pair			Momedu Pair		
White-White	0.012	0.007	College-College	0.013	0.008
White-Black	0.002	0.001	College-HS Grad	0.009	0.006
White-Hispanic	0.005	0.003	College-HS Dropout	0.004	0.003
White-Other	0.007	0.004	HS Grad-HS Grad	0.009	0.006
Black-Black	0.012	0.007	HSGrad-HS Dropout	0.006	0.004
Black-Hispanic	0.003	0.002	HS Dropout-HS Dropout	0.008	0.004
Black-Other	0.003	0.002			
Hispanic-Hispanic	0.005	0.003			
Hispanic-Other	0.003	0.002			
Other-Other	0.014	0.009			

Table 5A: Statistics Summary

Variable	Obs	Mean	Std. Dev.	Min	Max
Male	84654	0.501	0.500	0	1
Age	84365	14.985	1.713	10	19
White	84654	0.557	0.497	0	1
Black	84654	0.168	0.374	0	1
Hispanic	84654	0.169	0.375	0	1
Other	84654	0.106	0.307	0	1
Mom High School Dropout	84654	0.106	0.307	0	1
Mom High School Graduate	84654	0.445	0.497	0	1
Mom College Graduate	84654	0.254	0.436	0	1
Mom Education Missing	84654	0.195	0.396	0	1
School Size	84654	1016.033	599.820	44	2559
Cohort Size by Gender	83872	142.674	80.016	2	394
GPA	72276	2.806	0.805	1	4
Nominating Any Out-of-Grade within School Friend	84654	0.253	0.435	0	1
No. of People in Household	81496	4.291	1.143	1	6
No. of School Kids in Household	79019	0.712	0.928	0	6
Mother Born in US	74145	0.826	0.379	0	1
Father Born in US	61068	0.825	0.380	0	1
Born in US	81994	0.906	0.291	0	1
Live with Both Parents	80889	0.730	0.444	0	1
Live with Mother	81983	0.920	0.271	0	1
Live with Father	81691	0.763	0.425	0	1
Mother's Edu in Single Year	68137	13.371	2.366	0	17
Father's Edu in Single Year	53571	13.646	2.504	0	17
Adopt	81979	0.028	0.164	0	1
Health Condition at Birth	75910	0.019	0.135	0	1
No. of Valid Male Friend Nomination	84639	2.335	2.014	0	5
No. of Valid Female Friend Nomination	84639	2.418	2.056	0	5

Table 6A. First Stage—Correlation of Actual and Predicted Number of Friends

	Actual number of Friends						
	Total	College Grad	HS Dropout	Missing	Black	Hispanic	Other
Predicted Number of Friends							
Total	0.637**						
	(0.067)						
Mom College Grad		0.732**					
		(0.082)					
Mom HS Dropout			0.905**				
			(0.073)				
Mom Edu Missing				0.850**			
				(0.052)			
Black					0.670**		
					(0.077)		
Hispanic						0.757**	
						(0.106)	
Other							0.827**
							(0.085)
N	84654	84654	84654	84654	84654	84654	84654
R-sqared	0.222	0.198	0.143	0.115	0.345	0.209	0.231
F_iv	89.172	80.060	154.871	262.372	76.488	50.562	93.583
Fpvalue	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note: Predicted number of friends are generated from estimation using the pair level data and collapsed to student level. All regressions control for school-gender-student type fixed effect, school-grade fixed effect, and Guryan type control for school level friendship pattern omitting the student's contribution. Standard errors are clustered at the school level. **p<0.01 and *p<0.05.

Table 7A. Balancing Tests for Friend Choice Sorting with Student Demographic Characteristics

	Female		Male	
	College Grad Friends	HS Dropout Friends	College Grad Friends	HS Dropout Friends
Variables	(1)	(2)	(3)	(4)
Age	-0.00017	0.00059	-0.00046	-0.00003
	(0.00057)	(0.00049)	(0.00036)	(0.00032)
No. of People	0.00019	0.00002	0.00012	-0.00003
in Household	(0.00036)	(0.00027)	(0.00025)	(0.00020)
No. of School Kids	0.00008	0.00019	-0.00050	-0.00003
in Household	(0.00042)	(0.00025)	(0.00027)	(0.00017)
Live with Both	0.00143	0.00050	-0.00067	-0.00005
Parents	(0.00111)	(0.00053)	(0.00067)	(0.00044)
Live with Biological	-0.00007	-0.00201	0.00573	-0.00215
Parents	(0.00405)	(0.00306)	(0.00302)	(0.00314)
Mother's Edu in	0.00040	-0.00031	0.00027	0.00053
Single Year	(0.00037)	(0.00030)	(0.00044)	(0.00049)
Mother Born in US	-0.00051	0.00049	-0.00049	0.00122
	(0.00141)	(0.00098)	(0.00121)	(0.00085)
Born in US	0.00120	-0.00016	-0.00104	-0.00078
	(0.00183)	(0.00103)	(0.00135)	(0.00166)
Adopted	-0.00119	-0.00340	0.00161	0.00027
	(0.00290)	(0.00188)	(0.00189)	(0.00132)
Health Condition	-0.00232	-0.00057	-0.00144	-0.00002
at Birth	(0.00269)	(0.00113)	(0.00175)	(0.00140)
Observations	30,022	30,022	26,754	26,754
R-squared	0.973	0.960	0.981	0.946
Ftest	0.865	0.974	1.519	0.370
Fpvalue	0.567	0.469	0.139	0.958

Note: Each column displays a separate regression of the instrument variable--predicted number of mutual friends with college graduate mothers or with high school dropout mothers, on ten predetermined demographics variables. All regressions control for school-gender-student type fixed effect and school-grade fixed effect. Guryan type control for school level predicted number of mutual friends omitting the student him/herself is included in all regressions. Standard errors are clustered at the school level. **p<0.01 and *p<0.05.

Table 8A. Multivariate Robustness Test.

	Female GPA			Male GPA		
	2IV	3IV	5IV	2IV	3IV	5IV
No. of Friends with	0.194*	0.260**	0.250*	-0.005	-0.150	-0.155
Mom College Grad	(0.086)	(0.098)	(0.097)	(0.217)	(0.263)	(0.250)
No. of Friends with	0.013	0.058	0.052	0.063	-0.127	-0.129
Mom HS Dropout	(0.134)	(0.145)	(0.145)	(0.099)	(0.176)	(0.183)
Total No. of Friends		-0.069	-0.009		0.155	0.226
		(0.070)	(0.078)		(0.115)	(0.127)
No. of Black Friends			-0.144			-0.243
			(0.099)			(0.254)
No. of Hispanic Friends			-0.006			-0.137
			(0.143)			(0.230)
R-squared	0.042	0.026	0.030	0.023	0.028	0.022
N	36660	36660	36660	35116	35116	35116
Weak IV F-test statistic	38.000	23.938	15.352	14.670	11.195	5.698

Note: Each column displays a separate regression of GPA on number of mutual friends in different categories. Numbers of mutual friends are instrumented with corresponding predicted number of friends. All regressions control for school-gender-cross pair type fixed effect and school-grade fixed effect. Guryan type control for school level friendship pattern is included in both first and second stage. Standard errors are clustered at the school level. **p<0.01 and *p<0.05.

Table 9A. Factor Analysis Elements

	Survey Questions
Self Evaluation	How strong do you agree or disagree with each of the following statements? --I am physically fit. --I have a lot to be proud of. --I like myself just the way I am. --I feel like I am doing everything just right. --I have a lot of good qualities. In general, how hard do you try to do your school work well?
Environment Evaluation	How strong do you agree or disagree with each of the following statements? --I feel close to people at this school. --I feel like I am part of this school. --I am well coordinated. --The students at this school are prejudiced. --The teachers at this school treat students fairly. --I feel safe in my school. --I am happy to be at this school.
Mental Health	In the last month, --How often did you feel depressed or blue? --How often did you afraid of things? How strong do you agree or disagree with each of the following statements? --I feel loved and wanted. --I feel socially accepted.
Trouble at School	Since school started this year, how often have you had trouble: --getting along with your teachers? --paying attention in school? --getting your homework done? --getting along with other students?
Problematic Behavior	During the past twelve months, how often did you: --lie to your parents or guardians? --skip school without an excuse? In the past year, how often have you gotten into a physical fight?
Smoking and Drinking	During the past twelve months, --did you smoke cigarettes every week? --did you drink beer, wine, or liquor every week? --did you get drunk every week? Have you had a drink of beer, wine, or liquor—not just a sip or a taste of someone else's drink—more than two or three times in your life?
Health Status	In general, how is your health? How strongly do you agree or disagree with each of the following statements? --I seldom get sick. --When I do get sick, I get better quickly. In the last month, how often did a health or emotional problem cause you to: --miss a day of school? --miss a social or recreational activity?

Note: all variables from original dataset are converted to binary indicators to simplify the factor analysis.