

NBER WORKING PAPER SERIES

ESTIMATING THE EFFECTS OF FRIENDSHIP NETWORKS ON HEALTH BEHAVIORS
OF ADOLESCENTS

Jason M. Fletcher
Stephen L. Ross

Working Paper 18253
<http://www.nber.org/papers/w18253>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2012

We received valuable comments from numerous seminar participants at Baylor University, Cornell University, Lafayette College, Lehigh University, Texas A&M, University of California-Santa Barbara, University of Texas-Austin, Yale University, Population Association of American Conference, the Annual Health Econometrics Workshop, NBER Summer Institute, Urban Economics Association, and the Second Annual Economics of Risky Behaviors (AMERB) conference. We thank Yonghong An, Michael Anderson, Tao Chen, Ethan Cohen-Cole, Bill Evans, Don Kenkel, Brian Krauth, Anna Mueller, Bruce Sacerdote, Rusty Tchernis and Gautam Tripathi for specific comments that improved the paper. Fletcher and Ross gratefully acknowledge support from the NICHD (1R21 HD066230-01A1). Fletcher thanks the Robert Wood Johnson Foundation Health & Society Scholars program for its financial support.

This research uses data from Add Health, a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by a grant P01-HD31921 from the National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth@unc.edu). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2012 by Jason M. Fletcher and Stephen L. Ross. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Estimating the Effects of Friendship Networks on Health Behaviors of Adolescents

Jason M. Fletcher and Stephen L. Ross

NBER Working Paper No. 18253

July 2012

JEL No. I12

ABSTRACT

This paper estimates the effects of friends' health behaviors, smoking and drinking, on own health behaviors for adolescents while controlling for the effects of correlated unobservables between those friends. Specifically, the effect of friends' health behaviors is identified by comparing similar individuals who have the same friendship opportunities because they attend the same school and make similar friendship choices, under the assumption that the friendship choice reveals information about an individual's unobservables. We combine this identification strategy with a cross-cohort, within school design so that the model is identified based on across grade differences in the clustering of health behaviors within specific friendship patterns. Finally, we use the estimated information on correlated unobservables to examine longitudinal data on the on-set of health behaviors, where the opportunity for reverse causality should be minimal. Our estimates for both behavior and on-set are very robust to bias from correlated unobservables.

Jason M. Fletcher
Yale School of Public Health
Department of Health Policy and Management
60 College Street, #303
New Haven, CT 06520
and NBER
jason.fletcher@yale.edu

Stephen L. Ross
University of Connecticut
Department of Economics
341 Mansfield Road, Unit 1063
Storrs, CT 06269-1063
stephen.l.ross@uconn.edu

Estimating the Effects of Friendship Networks on Health Behaviors of Adolescents

Introduction

Individuals in modern societies are socially connected in a multitude of ways. For example, the social networking website Facebook.com has increased its membership by 100 million users in 2009, and now there are over 800 million users worldwide. Individuals use their social networks to receive and send information as well as establish, update, and enforce social norms of behavior. Both information acquisition as well as the impacts of social norms within social networks could have large effects on health behaviors, particularly adolescents, who are especially responsive to peer pressure (Brown et al. 1997). This heightening of peer influence also takes place during the developmental stage when many of the most costly health outcomes and behaviors are initiated. Our analysis will use detailed information on individual's health related behaviors and friendship networks from the National Longitudinal Study of Adolescent Health (Add Health) to examine the role of social interactions in these behaviors.

Many studies of social interactions find evidence of clustering of outcomes or behaviors above and beyond the clustering that might have been expected based on individuals' observables, including studies of crime (Glaeser, Sacerdote and Scheinkman 1996), employment (Topa 1999, Bayer, Ross, and Topa 2008), welfare usage (Bertrand, Luttmer, and Mullainathan 2000), pre-natal care (Aizer and Currie 2004), and youth health behaviors (Weinberg 2008).¹ We also observe unexpectedly high levels of clustering on health behavior within grades of students at the same school in our data. Specifically, if we look within schools, very little variation remains across grades in student composition in terms of racial or socio-economic variables, but we observe substantial across grade variation in health behaviors for student populations that are nearly identical. The purpose of this paper is to examine whether the within friendship clustering of health behaviors that lies underneath the clustering in specific grades is consistent with the influence of friendship networks.

Specifically, the primary purpose of our analysis is to examine the impact of friends' health behaviors on a student's own behavior while controlling for the likelihood that these students are friends because they have similar unobservables. Our controls for correlated unobservables are built on the idea that individuals who make the same friendship choices are likely to be more similar overall than might be indicated by their observables. Specifically, we

¹ See Ross (2011), Durlauf (2004) and Ioannides and Loury (2004) for relevant literature reviews.

examine a partial equilibrium model of friendship formation and use the model to illustrate the effect of controlling for fixed effects associated with clusters of observationally equivalent individuals who face the same friendship opportunity set and make the same “types” of friendship choices. We show that if individual students face an exogenous shock in terms of exposure to health behaviors, then as the number of friendship choices becomes large the unobservables of individuals in the same friendship choice cluster will converge to the same value and so a cluster fixed effect will act as a non-parametric control for unobservable attributes that influence friendship formation and might affect health behaviors. Further, we show that these controls reduce bias in the estimates of friendship effects for a small number of choices. Significantly, this structure also allows us to separate the influence of friends’ behaviors on individual behavior from the influence of the observable attributes of those friends, a key part of the reflection problem. This is true because the within friendship cluster comparisons are made between individuals who have observationally equivalent sets of friends and so have similar contextual effects, at least based on observables.

Our identification strategy relies on several empirical features of adolescent friendship networks. First, a large literature suggests that individuals exhibit strong racial, gender, and age preferences when choosing their friends—likes choose likes (Mayer and Puller 2008, Weinberg 2008). Second, data from the Add Health suggests that most friendships occur within grades, which is important for our use of cross-cohort variation in our identification strategy. Finally, as discussed above, individual grades within schools are quite homogenous over racial and socio-economic composition. Specifically, we will estimate models of youth drinking and smoking in high school that control for the share of same sex-same school-same grade friends who exhibit this behavior and fixed effects based on clusters of individuals who have the same race, ethnicity, and maternal educational attainment (individual observables), same school (same friendship opportunity set over observables), and same number of friends overall and for each racial, maternal education, and other demographic subgroups (similar friendship choices). Membership in a specific cohort in a school provides the shock in health behavior exposure, and in our preferred specification, we randomly choose one individual from each grade per cluster so that the model estimates are identified only by variation across cohorts within a school.

This strategy can be illustrated by the following thought exercise: consider a 9th grader and 10th grader who attend the same high school. As we show in detail below, these students face

very similar friendship opportunities with respect to racial, gender, and socioeconomic composition of their same-grade classmates, and yet there is substantial clustering of health behaviors into specific cohorts within schools. Thus, if we compare two students who choose similar “types” of friends based on race, maternal education, and other demographic characteristics, there will exist substantial differences in health behaviors between the across cohort friendship opportunities, and those differences in friends’ health behaviors is arguably quasi-random. The key is that the age difference between the 9th grader and the 10th grader (who attend the same high school and have the same preferences for “types” of friends) has effectively randomized these two students into their actual friendship network.

We find evidence that this strategy produces somewhat smaller (no more than 10% smaller) “network effect” estimates than the more standard school fixed effect models, and after controlling for correlated unobservables we still find very robust evidence of network effects on the smoking and drinking behavior of adolescents. Further, we find that peer health behaviors tend to be weak predictors of predetermined student or family attributes with controlling for friendship choices reducing estimated “effects” on predetermined attributes by between 45 and 97% relative to school fixed effect estimates. In addition, we demonstrate that the effect of controlling for friendship choices on estimates of the influence of friends is quite similar across subsamples of students with different numbers of friends. We find no evidence of falling causal estimates of friends’ behaviors as we increase the number of friends, suggesting that the large number of friendship demographics included in the construction of the fixed effect are sufficient to control for individual unobservables even with a small number of actual reported friends.

Finally, while our fixed effect estimator insulates us from bias associated with correlated unobservables, neither the fixed effects nor our counterfactuals address concerns about reverse causation where for example individuals who smoke choose friends who smoke. In a second analysis, we show how our controls for correlated unobservables from the cross-sectional analysis can be combined with longitudinal data to examine the effect of friends’ smoking or drinking behavior at baseline on the onset of smoking or drinking prior to follow-up interviews one year later. As in cross-sectional analyses, we find robust evidence that having friends who smoke or drink contributes to the onset of smoking or drinking, and the inclusion of our controls for unobservable student attributes have little impact on those estimates.

Background Literature

A large body of research across multiple disciplines has shown very strong correlations in health behaviors for individuals who are socially connected. One reason there has been so much research and policy interest in exploring how networks affect health behaviors and outcomes is the potentially large set of health interventions and policies that could be proposed to leverage social influences on health behaviors. While the promise of using social networks to affect health is compelling, so too are the empirical issues inherent in detecting causal effects of social networks using observational data.

Social interactions between individuals are believed to matter for those individuals' outcomes across a variety of dimensions including labor market success, criminal activity, educational performance and health related behaviors. However, evidence that social interactions between friends, neighbors or acquaintances actually influence outcomes is usually quite indirect. For example, the recent Moving To Opportunity field study (Kling, Liebman, Katz, 2007) demonstrates that moving to a lower poverty rate neighborhood improves the mental health outcomes of mothers and the academic outcomes of their female children, but we do not know whether these improvements arise due to exposure to broad environmental factors or because the change in environment influences the individuals with whom these women and children regularly interact socially. Similarly, a growing literature on peer effects in schools (Gould, Lavy, and Passerman 2004, Lavy and Schlosser 2007, Lavy, Passerman and Schlosser 2008, Friesen and Krauth 2008, Bifulco, Fletcher and Ross 2011) documents a causal effect of school demographics, such as gender, race, student ability and parental education, on educational and health related outcomes using apparently random variation across cohorts of students at the same school. Again, however, it is unknown whether these peer effects arise because cohort composition affects the educational environment or because cohort composition affects the demographic composition of each student's friends and acquaintances.

A few studies have provided evidence that explicitly suggest a causal effect of social interactions on outcomes. For example, Bertrand, Luttmer and Mullainathan (2000) show that individuals who reside near others who speak the same language are more likely to use welfare if they belong to a language group with high welfare usage suggesting that communication along social links can contribute to welfare usage. Bayer, Ross and Topa (2008) find that people who live on the same block and share key attributes in common, such as having children of similar

age, are much more likely to work in the same location, and suggest that this cluster arises from labor market referrals. Bayer, Hjalmarsson and Pozen (2009) show that sharing time in prison with other juvenile offenders who have experience with a particular crime leads to a higher likelihood of an offender committing that crime in the future.

However, without exploiting information on actual social interactions between individuals, it is very difficult to know whether these clusters of outcomes are driven by social interactions or possibly heterogeneous effects of shared environment. For example, Aizer and Currie (2004) find a strong correlation within groups in the likelihood of expectant mothers using pre-natal care, but then demonstrate that these effects cannot arise due to information sharing between mothers because the estimated effects are the same size for mothers who have previously used the program. Further, recent work on peer effects in schools points to classroom disruption and the impact of these disruptions on the classroom environment as a mechanism (Lavy and Schlosser 2007, Aizer 2009, Hoekstra and Carrell In Press),² rather than through social relationships between students.

The estimation of the causal effects of observed social interactions between individuals is particularly challenging (Manski 1993). First, individuals who are friends share common unobservables because they chose each other as friends, because they self-select into the same social network, and because individuals in the same social network are simultaneously affected by their shared environment. Second, it is difficult to separate the influence of an individual's behavior and an individual's attributes in determining the health behaviors of his or her friends. Third, individuals may select their friends or peers based on the behavior of those peers, rather than adapting their behavior to the behavior of friends whom they have already selected. Unfortunately, failure to overcome these empirical difficulties casts considerable doubt on the current knowledge base linking the health behaviors among individuals in the same social network. Providing evidence of the causal mechanisms and the likely effects of policies is essential to be able to properly leverage social network effects on health behaviors.

A few studies have attempted to identify the causal effect of observed social relationships through a variety of mechanisms. Christakis and Fowler (2007, 2008) examine the effect of

² Lavy and Schlosser (2007) find that gender composition affects both academic outcomes and the levels of classroom disruption. Aizer (2009) and Hoekstra and Carrell (In Press) find negative effects of having peers with ADHD or peers who experience domestic abuse at home and so are more disruptive in the classroom, and in both cases these effect weaken when either the ADHD is diagnosed or the domestic abuse is reported.

social networks on obesity and smoking in the medical literature arguing that the inclusion of lags of the outcome for both the respondent and his/her friend are sufficient to control for common environmental factors. Renna et al. (2008) and Trodgon et al. (2008) also examine social contagion in obesity, but control for the endogeneity of friendship using a combination of school fixed effects and instrumental variables for friend's obesity (obesity status of friend's parents and in the case of Trodgon et al. also own birth weight and parent's health status). Halliday and Kwak (2009) examine the effect of friends' obesity controlling for school and individual fixed effects. Calvó-Armengol et al. (2009) and Patachini and Zenou (In Press) examine the effect of friendship links on educational outcomes and crime, respectively, controlling for network fixed effects and, in Patachini and Zenou, instrumenting for friend's criminal behavior using the criminal behavior of the friends own self-reported friends.

The identification strategies pursued in all of these papers have significant limitations and in several cases have been questioned in the literature.³ Cohen-Cole and Fletcher (2008a) show that adding controls for environmental factors eliminates the social network effects for obesity detected in models that use outcome lags as controls for unobserved individual and environmental factors.⁴ Halliday and Kwak (2009) show that contagion effects identified in models that control for school fixed effects are not robust to the inclusion of individual fixed effects. Further, while network fixed effects control for within school selection into social networks, the formation of human social networks tends to generate one large connected network that usually includes a majority of the individuals in the network structure and the vast majority of the social links in the network structure, and so most of the within school variation in friend's behaviors is retained in a within network analysis (Jackson 2008, Ross 2011). Finally, instrumenting for friend's behaviors with either parental attributes or the behavior of the friend's own self-reported friends may be problematic because these variables or correlates of these variables are likely to be observable when friendships are formed.

In the next section, we propose a new identification strategy arguing that observationally similar individuals who make similar friendship choices are likely to be similar on unobservables, and so identify the effect of friends' behavior by comparing individuals who

³ One exception is Nair et al. (2006) who find peer effects in prescription decisions using data on doctors' prescriptions and the physicians on whose opinions they rely. They use the release of new drug guidelines as an exogenous shock to peer prescription behavior

⁴ Cohen-Cole and Fletcher (2008b) show more generally that parsimonious models that control for lagged outcomes will produce social network effects even in outcomes where none are expected to exist, such as for height.

made similar friendship choices but have friends that exhibit different behaviors. We develop a simple model of friendship formation and demonstrate circumstances under which consistent estimates of the effect of friends' health behavior on own health behavior can be uncovered using this strategy, and show that identification in the presence of correlated unobservables requires an exogenous shock in exposure to potential friends who exhibit certain behaviors. Following the literature on peer effects, we propose that across cohort variation within schools can provide this exogenous variation in exposure to health behaviors and demonstrate empirically that health behaviors vary substantially more across cohorts than student attributes, like race or parental education, evidence consistent with our identification strategy.⁵

This approach is similar to earlier analyses by Dale and Krueger (2002) and Fu and Ross (In Press) who use fixed effects for individuals who are equivalent on key attributes (college application and admission decisions or residential location choice) and then have the same outcome or make the same choice as a reduced form control in order to minimize bias from unobservables. However, our analysis has the advantage over these earlier studies because the identification strategy contains a clear source of exogenous variation that can create within cluster differences in environment, namely differences in exposure to health behaviors associated with belonging to a particular cohort or grade of students. Further, our friendship formation model demonstrates the importance of having such a source of exogenous variation for identification when attempting to identify causal effects in models with endogenous social interactions.

Identification Strategy

In this paper, we seek to estimate the effect of friends' health behaviors while overcoming several of the key empirical obstacles that we outlined above. The primary focus of our analysis is to address selection into friendships and other social relationships based on a student's unobservables, but our fixed effect strategy also addresses reflection, at least on observables, in that students in the same cluster have friends with the same observables. In

⁵ Two studies combine across cohort variation with an aggregation strategy in order to identify the effect of friendship. Babcock (2008) finds that being part of a cohort of more socially connected students raises the likelihood of high school graduation and college attendance, and Nathan (2008) finds that racial heterogeneity of friendships improves academic performance especially among whites. However, as with previous studies that exploit aggregation, such as Culter and Glaeser (1997) on segregation or Evans et al. (1992) on neighborhood poverty, it is impossible to know if the effects are driven by the implied social connections between individuals or the associated aggregate environment (Ross 2011).

addition, in our follow-up analysis, we examine the on-set of health behaviors where reverse causality is unlikely to be a concern.

The intuition behind our approach is that we seek to form comparison groups based on information in the data that describes the friendship *options* of students as well as the students' *choices* of friends (given these options) following the premise that individuals who make similar decisions or have similar outcomes when facing the same set of options likely are very similar on both observable and unobservable attributes. The beginning of this section illustrates this intuition, the next two subsections derive formal results.

We begin with a slight modification to the relatively straightforward linear-in-means model of social interactions (Case and Katz 1991, Manski, 1993; Moffit, 2001; Brock and Durlauf, 2001) by restricting social interactions to arise from a subset of individuals "friends" within a social environment (or school s) and dividing the unobservable into two components: an unobservable that also affects friendship choice ε_i and an orthogonal unobservable error that does not enter the friendship choice model μ_i .⁶ Specifically, we consider the following empirical model:

$$H_{is} = \beta_s + \left(\frac{1}{n_i} \sum_{j \in \Omega_{is}} H_{js} \right) \beta_1 + \left(\frac{1}{n_i} \sum_{j \in \Omega_{is}} X_j \right) \beta_2 + X_i \beta_3 + \varepsilon_{is} + \mu_{is} \quad (1)$$

where H_{is} indicates a particular health behavior, such as smoking, of individual i in a broad social environment or school s , X_i contains the individual's observable attributes, n_i is the number of friends of person i or more generally the number of friendship choices, Ω_{is} defines the set of individual i 's friends in s , H_{js} and X_{js} indicate the health behavior and observable attributes of individuals within Ω_{is} , and β_s is a school fixed effect.

As Manski (1993) demonstrates, even without the correlations in social networks that are caused by sorting into and within networks based on unobservables, e.g. ε_{is} orthogonal to

⁶ An alternative specification might involve a single unobservable each for determining health behavior and friendship outcomes. The specification in equation (1) is equivalent to such a model with the imposition of one restriction. We start with a model where the composite unobservables in equation (1) and a friendship formation model, $\tilde{\mu}_{is}$ and ε_{is} , are correlated, and then we can define μ_{is} as $\tilde{\mu}_{is} - E[\tilde{\mu}_{is} | \varepsilon_{is}]$ where we assume that the $E[\tilde{\mu}_{is} | \varepsilon_{is}] = \alpha_0 + \alpha_1 \varepsilon_{is}$ so that the composite error $\tilde{\mu}_{is}$ depends upon the uncorrelated disturbances μ_{is} and ε_{is} and α_1 is simply initialized to one in the health behavior model and generality is maintained by allowing ε_{is} to enter the friendship formation model in a general manner.

$\left(\frac{1}{n_i} \sum_{j \in \Omega_i} X_j\right)$, this model is intrinsically unidentified.⁷ This occurs because group member characteristics that might explain the health of group members j and so act as instruments for health behavior cannot be excluded from the second stage regression for the health behaviors of i because these attributes may just as reasonably directly influence i 's behaviors (the reflection problem).^{8 9}

Our identification strategy is to sort students into clusters c based on comparing similar students who faced similar friendship options and made similar friendship choices. This sorting is based on both observable (to the researcher) and unobservable characteristics. Following the standard selection argument: if two individuals make similar choices and differ on observables, then they are expected to differ on unobservables, as well (Heckman, 1976). Similarly, if two individuals are the same on observables and make similar choices, they are expected to be quite similar on unobservables (Altonji, Elder, and Taber 2005). Therefore, as argued by Dale and Krueger (2002) and Fu and Ross (2010), the inclusion of fixed effects for such clusters should assure that we are comparing students who are similar on both observables and unobservables, which weakens the correlation between peers' behaviors and a student's unobservable characteristics. Further, since all students in a cluster should have similar observable characteristics, the inclusion of the fixed effect also captures the observables associated with the students' peers while allowing the effect of behavioral differences within a cluster to identify the effect of friends' behavior on individual behavior. This feature of the approach solves the empirical problem outlined above and isolates the causal effect of student behaviors on the behavior of their friends from the effect of observable friends' attributes.

⁷ By this we mean that there is insufficient information in the regression to estimate uniquely the parameters of interest (β_1 in particular).

⁸ For example, if one observes clustering of criminal behavior among friends whose parents have less education, even after controlling for all possible individual and environmental factors that might explain such clustering available in the data, we still cannot conclusively determine whether the clustering is caused because having friends whose parents have less education contributes to criminal behavior or individuals whose parents have less education are more likely to engage in criminal behavior and such criminal behavior influences the behavior of the individual's friends. See Brock and Durlauf (2001, 2006) for recent methodological progress on this problem.

⁹ As noted by Sacerdote (2001) and Bayer and Ross (2008), when social network effects are determined in part by unobservable characteristics, even random assignment cannot solve this identification problem. While random assignment breaks the correlation between the health behavior i 's peers and i 's unobservable characteristics, the coefficient estimate on the behavior of peers is a composite of both the direct effect of peer's behaviors and the effect of peers' unobservable characteristics.

Specifically, define a cluster of individuals c in the same school who are observationally equivalent on X_i and choose observationally equivalent friends based on X_j . This structure implies that the individual and friendship group observables are the same within a cluster so that the contribution of the variables that determine clusters to individual's health behavior are constant within cluster or

$$\beta_s + \left(\frac{1}{n_i} \sum_{j \in \Omega_s} X_j \right) \beta_2 + X_i \beta_3 = \beta_s + \left(\frac{1}{n_i} \sum_{j \in \Omega_{ks}} X_j \right) \beta_2 + X_k \beta_3 \quad (2)$$

for all $i, k \in c$. Further, we assume that the model that defines selection over friendships on health behaviors and on observable attributes depend monotonically on the same observable vector of attributes X_i and the same single index unobservable ε_{is} . This assumption is central to our identification strategy. Without monotonicity, multiple values of the unobservable might be consistent with the observed friendship choices for observationally equivalent individuals. With monotonicity on the unobservable, similar individuals who face the same type of friendship options based on the available social network (s) and make similar choices should have similar values on their unobservable because if they differed substantially on the unobservable they would likely have made different friendship choices.

Specifically, we can define ρ_c as a cluster fixed effect where based on the discussion in the preceding paragraph

$$\rho_c \approx \beta_s + \left(\frac{1}{n_i} \sum_{j \in \Omega_s} X_j \right) \beta_2 + X_i \beta_3 + \varepsilon_{is} \approx \beta_s + \left(\frac{1}{n_i} \sum_{j \in \Omega_{ks}} X_j \right) \beta_2 + X_k \beta_3 + \varepsilon_{ks} \quad (3)$$

Further, based on the construction of μ as an idiosyncratic disturbance, $E[\mu_{is} | \rho_{c_i}] = 0$ and substituting equation (2) into equation (1) yields

$$H_{ics} \approx \left(\frac{1}{n_i} \sum_{j \in \Omega_s} H_j \right) \beta_1 + \rho_c + (\mu_{is} - \bar{\mu}_c) \quad (4)$$

where $(\mu_{is} - \bar{\mu}_c)$ represents the deviation of the right hand side expression in equation (3) from the average of this expression for all individuals in cluster c , $\bar{\mu}_c$.

Next, in order to understand the circumstances when our fixed effect estimator will yield consistent estimates of the effects of friendship networks, we develop a partial equilibrium model

of friendship formation and use the friendship model to examine the properties of the specification in equation (4).

Partial Equilibrium Model of Friendship Formation

We begin this subsection by repeating equation (1)

$$H_{is} = \beta_s + \tilde{H}_{is}\beta_1 + \tilde{X}_{is}\beta_2 + X_i\beta_3 + \varepsilon_{is} + \mu_{is} \quad (5)$$

where we define \tilde{H}_{is} and \tilde{X}_{is} as $\frac{1}{n_i} \sum_{j \in \Omega_{is}} H_{js}$ and $\frac{1}{n_i} \sum_{j \in \Omega_{is}} X_j$, respectively, restricting H_{is} to only

take on the values of 1 (healthy) or 0 (unhealthy) and X_i to only take on the values 1 (good) or 0 (bad) where the good type is defined agnostically as the type that is more likely to exhibit healthy behavior, and without loss of generality assume that β_2 and β_3 are non-negative.¹⁰ Further, we assume that μ_{is} is an idiosyncratic error so that

Assumption 1: $E[\mu_{is} | \tilde{H}_{is}, \tilde{X}_{is}, X_i] = 0$

Now we define the likelihood of observing a specific health behavior H_{is} and type X_i for a selected friend by the following general set of functions

$$Pr_{is}[X_j = x, H_j = h | X_i, \varepsilon_{is}, \pi_{is}, j \in K_i] = f_{sxh}(X_i, \varepsilon_{is}, \pi_{is}) \quad (6)$$

where π_{is} is an additional unobservable that does not enter equation (5), but influences friendship formation over health behaviors. The function f_{sxh} is defined over the four combinations of the outcomes for X and H and can vary across schools s since the social environment varies across schools. The four probabilities must sum to one for a given school for any value of the functions' arguments.

We assume that the probabilities of having a friend who is of good type and who exhibits healthy behavior are not directly influenced by own health behavior (Assumption 2), are monotonic in the individual's unobservable attributes that influence health behavior (Assumption 3), and that additional unobservable attributes exist that have a monotonic influence on friendship formation concerning health behavior, but have no influence on either own health behavior or friendship formation over other friendship attributes (Assumption 4). While the unobservables might be correlated with X_i , some variation must remain in the unobservables

¹⁰ See Brock and Durlauf (2001, 2006) for an alternative identification approach for the reflection problem that applies when behavior is discrete.

associated with forming friends with specific health behaviors that does not enter own health behavior after conditioning on X_i . These assumptions can be summarized as follows

Assumption 2: $\frac{\partial f_{s11}}{\partial H} = 0, \frac{\partial f_{s10}}{\partial H} = 0, \frac{\partial f_{s01}}{\partial H} = 0, \frac{\partial f_{s00}}{\partial H} = 0.$

Assumption 3: $\frac{\partial f_{s11}}{\partial \varepsilon} + \frac{\partial f_{s10}}{\partial \varepsilon} > 0$ and $\frac{\partial f_{s11}}{\partial \varepsilon} + \frac{\partial f_{s01}}{\partial \varepsilon} > 0.$ ¹¹

Assumption 4: $\frac{\partial f_{s11}}{\partial \pi} = -\frac{\partial f_{s10}}{\partial \pi} > 0, \frac{\partial f_{s01}}{\partial \pi} = -\frac{\partial f_{s00}}{\partial \pi} > 0,$ and $Var[\pi_{is} | X_i] \neq 0$

While Assumption 3 will be maintained throughout, we will examine the implications of relaxing Assumption 2 in the next subsection by allowing own health behavior to influence friendship formation over friends' health behavior. Assumption 4 is designed to capture the across cohort variation described in our identification strategy. Our maintained assumption is that membership in a cohort is based on age and so exogenous conditional on school, and so is not directly associated with own health behavior, except of course through the well-known age-gradient in unhealthy behaviors such as smoking and drinking. Further, cohort membership creates a shock to the health behavior composition of potential friends while leaving the exogenous attributes of potential friends relatively unchanged. The implications of violating assumption 4 are further examined in the appendix.

Now, we define a cluster c as all students in a school who are of the same type, have the same number of friends, and make the same friendship choices over “friendship type”.

Definition 1: A cluster c in school s is defined so that $X_{is} = X_{ks}, n_i = n_k$ and

$$\frac{1}{n_i} \sum_{j \in \Omega_{is}} X_{js} = \frac{1}{n_k} \sum_{j \in \Omega_{ks}} X_{js} \text{ for all } i \text{ and } k \text{ in cluster } c \text{ and there exist no individuals } l \text{ outside of}$$

$$\text{cluster } c \text{ where } X_{is} = X_{ls}, n_i = n_l \text{ and } \frac{1}{n_i} \sum_{j \in \Omega_{is}} X_{js} = \frac{1}{n_l} \sum_{j \in \Omega_{ls}} X_{js}.$$

Our first important result is that the bias in our estimate of β_1 in equation (4) limits to zero as the number of friends, or more generally the number of friendship choices, becomes large. In our empirical work, the number of choices made is substantially larger than the number

¹¹ The assumption of a positive relationship between good type and the individual's friendship formation propensity y_{is} is made without loss of generality because one can reverse the relationship by designating healthy behavior as unhealthy. However, once this assumption is made, the sign of the relationship between y_{is} and having friends who exhibit healthy behavior is meaningful. If this relationship is positive, then one's type has the same effect on health behavior composition of friendships as it has on composition of friends over type, and this assumption cannot be undone by reversal because the definition of what individual type means is nailed down by β_3 and the coefficient of one on ε_{is} in equation (5)

of friends because students choose over many demographic attributes including race, ethnicity, and maternal education. Further, later in the paper, we demonstrate that our results are robust to focusing on the subsample of students with the largest number of friends.¹²

Proposition 1: Under Assumptions 1 through 4 plus Definition 1, the bias arising from estimating the cluster fixed effects model in equation (4) limits to zero as n_i becomes large for all i in the sample.

Proof: See Appendix

A key derivation in Proposition 1 is that the bias in the cluster fixed effects estimate of β_1 can be written as

$$\phi_1^c = \frac{\text{Cov}[\varepsilon_{is} - \bar{\varepsilon}_c, (\tilde{H}_{is} - \bar{H}_c)]}{\text{Var}[(\tilde{H}_{is} - \bar{H}_c)]} \quad (7)$$

where $\bar{\varepsilon}_c$ and \bar{H}_c are the cluster specific mean of ε_{is} and \tilde{H}_{is} , respectively. The bias limits to zero because, as the number of choices becomes large, two individuals i and k can only belong to the same cluster if $\varepsilon_{is} = \varepsilon_{ks}$. However, the within cluster variation in ε cannot limit to zero while within-cluster variance of \tilde{H}_{is} remains unless \tilde{H}_{is} contains variation associated with π . Therefore, the non-zero variance assumption in Assumption 4 is crucial to the consistency of our estimator.

Second, even when the number of friends is small, we can show that the inclusion of cluster fixed effects reduces the bias in estimates of the effect of friend's health behavior on own health behavior with the imposition of a couple of additional assumptions. First, we create a linear projection of \tilde{H}_{is}

$$\tilde{H}_{is} = \lambda_s + \tilde{X}_{is}\lambda_1 + X_i\lambda_2 + V_{is} \quad (8)$$

such that $V_{is} = V(\tilde{X}_{is}, X_i, \varepsilon_{is}, \pi_{is})$. We assume that the conditional expectation of V_{is} is zero and that the conditional variance of V_{is} is less than or equal to the variance of V_{is} .

Assumption 5: $E[V_{is} | \tilde{X}_{is}, X_i] = 0$ and $\text{Var}[V_{is} | \delta_c] \leq \text{Var}[V_{is}]$.

¹² A second implicit assumption is that the number of observations or students increases more quickly than the number of friendship choices. This assumption is required in order to assure consistency in the fixed effect estimates. Later, in our empirical work, we demonstrate that key findings on behavior on-set are robust in subsamples that focus on individuals in friendship choice clusters that contain a larger number of students.

The first part of Assumption 5 implies that

$$E[\tilde{H}_{is} | \tilde{X}_{is}, X_i] = \lambda_s + \tilde{X}_{is}\lambda_1 + X_i\lambda_2 \quad (9)$$

This restriction is essentially a law of large numbers style assumption where we assume that the average of this residual is zero over repeated realizations of \tilde{H}_{is} and \tilde{X}_{is} for a given X_i . This assumption would be standard if \tilde{X}_{is} did not depend upon ε_{is} . The second half of Assumption 5 is something that can be theoretically violated in principle, but in practice we expect that variances will decline after conditioning on additional information, and we also directly verify this assumption for estimates of V_{is} using our data.

Proposition 2: Under Assumptions 1 through 5 plus Definition 1, the bias arising from estimating the cluster fixed effects model in equation (4) has the same sign and is smaller than the bias that arises for the OLS model described in equation (5).

Proof: See Appendix

In the appendix, we also consider a situation where the shock to friendship formation over health behavior π_{is} also affects friendship formation over the exogenous attributes, which might be the case if we were considering two individual in the same grade who had unobservable differences (associated with themselves rather than the cohort) that lead them to choose friends with different health behaviors and so likely also caused them to choose friends with different attributes. We show that Proposition 1 may not hold when we relax Assumption 4 along this margin, and in fact the sign of the bias may be reversed relative to OLS when we control for friendship cluster FE's that contain variation that does not satisfy Assumption 4, such as within grade variation in friends' health behaviors.

Simultaneity of Health Behavior and Friendship Sorting Model

In this section, we extend the friendship formation function so that friendship formation over health behavior depends upon one's own health behaviors, creating true simultaneity between one's own health choices and the selection of friends based on their health choices. Specifically, we relax Assumption 1 so that own health behavior influences the likelihood of having friends who exhibit a health behavior, but do not allow own health behavior to affect friendship formation over the observable attributes. So

$$Pr_{is}[X_j = x, H_j = h | H_i, X_i, \varepsilon_{is}, \pi_{is}, j \in K_i] = f_{sxh}(H_i, X_i, \varepsilon_{is}, \pi_{is}) \quad (10)$$

with

Assumption 6: $\frac{\partial f_{s11}}{\partial H} = -\frac{\partial f_{s10}}{\partial H} > 0$ and $\frac{\partial f_{s01}}{\partial H} = -\frac{\partial f_{s00}}{\partial H} > 0$.

Therefore, the idiosyncratic error μ_{is} does not have a conditional expectation of zero because it influences the health behavior of friends \tilde{H}_j through one's own health behavior, and the bias in the coefficient on friend's health behaviors contains a second term φ_1 .

Proposition 3: Under Assumptions 1 and 3 through 6 plus Definition 1 the cluster fixed effects model estimate of the effect of friends' health behavior limits to a reduced form estimate that is the sum of β_1 and a second term that captures the correlation between μ_{is} and \tilde{H}_{is} (φ_1) as n_i becomes large for all i in the sample.

Proof: See Appendix.

Proposition 3 illustrates why our fixed effects estimator only addresses bias from correlated unobservables, but not endogeneity, where smoking causes someone to select friends who smoke. If β_1 is zero, it is still possible for φ_1 to be non-zero, because high values of μ_{is} will increase the likelihood of smoking and as a result lead to a larger numbers of smoking friends other things equal, V_{is} . We cannot address this limitation theoretically, but later in the paper we examine an empirical model on the onset of health behaviors where reverse causality should not be a serious concern. At that time, we will present the methods for conducting the required two stage estimates.

Friendship Data

In order to accomplish our research goals, we use the only available national dataset containing rich friendship network information as well as health behaviors, the National Longitudinal Study of Adolescent Health (Add Health). The Add Health is a school-based, longitudinal study of the health-related behaviors of adolescents and their outcomes in young adulthood.¹³ For this paper, we focus on the In-School data collection, which utilized a self-

¹³ In short, the study contains an in-school questionnaire administered to a nationally representative sample of students in grades 7 through 12 in 1994-95 and three in-home surveys that focus on a subsample of students in 1995 (Wave 1), and approximately one year later (Wave 2) and then six years later (Wave 3). The fourth wave of the survey was collected in 2008/9. The study began by using a clustered sampling design to ensure that the 80 high schools and 52 middle schools selected were representative of US schools with respect to region of country, urbanicity, size, type, and ethnicity. Eligible high schools included an 11th grade and enrolled more than 30

administered instrument to more than 90,000 students in grades 7 through 12 in a 45- to 60-minute class period between September 1994 and April 1995. The questionnaire focused on topics including socio-demographic characteristics, family background, health status, risk behaviors, and friendship nominations. In particular, each student respondent was asked to identify up to 10 friends (5 males, 5 females) from the school's roster. Based on these nominations, social networks within each school can be constructed and characterized, linking the health behaviors of socially connected individuals.

Of the nearly 90,000 students in the schools originally surveyed, several reductions in the sample size were made in order to construct the analysis sample. First, nearly 4,500 students did not have individual identification numbers assigned. Nearly 12,000 students did not nominate any friends and 5,000 individuals nominated friends who were not able to be linked with other respondents due to nominations based on incomplete information ("nicknames" rather than names, or the nominated friend did not appear on the Add Heath school roster, etc.) These issues reduced the sample to approximately 66,000 respondents. Appendix Table 1A presents an analysis of the correlates associated with individuals being dropped from the sample for these reason discussed above, as well as additional sources of selection arising from the empirical specification discussed below.¹⁴

In this paper, our main focus is on individuals with same-sex/same-grade level friends, which reduces the sample to approximately 55,000 students.¹⁵ One reason to focus on same-sex friends is that romantic relationships may be nominated as "friends". In addition, most previous studies of friendship networks also limit the network definition to same-sex friends. We limit our analysis to same-grade friends in order to use cross-cohort (grade) variation in friendship opportunities and choices, as we describe below. In order to retain sample size, we impute

students. More than 70 percent of the originally sampled high schools participated. Each school that declined to participate was replaced by a school within the stratum.

¹⁴Briefly, race, gender, family structure, and missingness on other variables predicts sample selection in to the original 66,000 observations to some extent, however health behaviors are not robust important predictors. In regards to same-sex/same-grade friendship nominations, the likelihood of making such nominations increases by grade and is smaller for more advantaged students. We find that the proportion of smokers in the grade (potential friends) is not related to these nomination patterns, however, individuals with drinking grademates are slightly more likely to nominate same-grade/same-gender friends (a 10 point increase in grademates drinking is associated with a 1 percentage point increase in the probability).

¹⁵ Of the 66,000 students, 4,300 do not nominate any same grade friends and 4,100 do not nominate any same-grade/same-gender friends (that is, they nominate same grade friends but no same-grade/same gender friends).

missing covariates, such as maternal education, and control for missingness, but we do not impute missing outcomes.

Table 1 presents descriptive statistics of the analysis sample and shows that approximately 34% of the sample reports smoking and 54% of the sample reports drinking alcohol. The average adolescent nominates 2.8 same-sex friends. In Table 2 we present the distribution of friends' health behaviors in the data. Friendship networks include considerable variation, including individuals who have no smoking/drinking friends through individuals who have all smoking/drinking friends.

Evidence of Variation in Friendship Options

As we demonstrate above, identification of the effect of friend's health behavior requires a shock in exposure to potential friends with specific health behaviors. In our empirical analysis, we control for fixed effects associated with similar students who make similar friendship choices on student attributes, but because they belong to different cohorts of the same school draw groups of friends who exhibit differing health behavior. That is, the dataset contains multiple cohorts within each surveyed high school, which allows us to combine our friendship type fixed effects with the use of cross-cohort, within-school variation and in doing so are able to compare students who face similar friendship options (are in the same school) and make similar friendship choices. This extension relies heavily on the assumption that individuals who attend the same school, but different grades, have essentially the same "types" of friendship options.

To what extent do students in the same school face similar friendship options? Using the Add Health data, we show below in Table 3 that controlling for school and grade effects can predict over 95% of the variation in racial composition of potential friends (classmates) in the data. Likewise, controlling for school and grade predicts 93% of the variation in peers' maternal education level and 96% of the variation in classmate nativity. These findings suggest that students in different grades but who attend the same school have very similar friendship options based on race and family background of peers.

In addition, there is substantially more variation across cohort, within schools in unhealthy behaviors. Using the same regression analysis, our data show that we only predict 77% of peer smoking rates and 81% of peer drinking rates. Thus, these results suggest that there is substantial variation in exposure to health behaviors of potential friends (classmates) even within

school, while at the same time the friendship options based on race, maternal education, and nativity is nearly identical for students across grades within the same school.

Empirical Specification

Our friendship clusters are based on students in the same school choosing sets of friends with very similar demographic attributes. As there is evidence that adolescents have strong preferences to befriend classmates based on age, gender, and race (Mayer and Puller 2008; Weinberg 2008), we create our “individual type-friendship type clusters” by focusing primarily on those attributes. Given a limited sample, there is clearly a trade-off between how restrictive we make our definitions of observationally similar individuals and of same friendship types. We begin by placing the most weight on obtaining very specific “friendship-type” clusters because this is dimension over which students are making choices.

The friendship clusters are based on the following exogenous characteristics of chosen friends, including (1) race (black vs. Hispanic vs. white vs. Asian vs. other) (2) maternal education (no college vs. some college vs. college graduate) (3) family structure (living with mother vs. not living with mother) and (4) nativity (native vs. foreign born). Specifically, the number of friends chosen from a group for each characteristic is used in the cluster. Importantly, our clusters are quite flexibly created, such that an individual who chooses five black friends is in a different cluster than an individual who chooses four black friends.¹⁶

We define observationally equivalent students over two key variables: the student’s race (white, black, Hispanic, and Asian) and whether their mother is a college graduate for the creation of individual type-friendship type clusters. This more parsimonious specification of observationally equivalent students has the added advantage of allowing for additional counterfactuals over those student observables. For family structure and nativity, we can examine whether within cluster differences in friend’s drinking or smoking can explain whether the student lives with their mother or whether the student is an immigrant. Similarly, for years of maternal education, we can test whether within cluster variation (restricting comparisons based on whether their mother completed college or not) in friends’ behavior explain differences in

¹⁶ As an example, friendship cluster 15 could be created based on nominating four friends such that: friend A is white, has a college educated mother, lives with his mother, and is native born; friend B is white, has a mother with some college, lives with his mother, and is native born; friend C is white, has a college educated mother, lives with his mother, and is foreign born; friend D is black, has a college educated mother, lives with mother, and is native born. Cluster 16 could be identical except the individual nominated four white friends instead of three white friends and one black friend; Cluster 17 could be identical to cluster 15 except all the nominated friends are native born.

students' own maternal education. Finally, we will present counterfactuals for paternal attributes, education and nativity that are not used in either the definition of individual type or friendship choices.

In our final and preferred model, as discussed above, we restrict our comparisons to students in different grades who are observationally equivalent on race and maternal education and chose the same friendship set on those attributes plus nativity and living with mother.¹⁷ Friends' health behaviors are based on own-grade friendships, and so these estimates are based entirely on comparisons across cohorts. Specifically, one student's friends' health behavior could not vary from another student's in the comparison group because one student selected a given student and another selected away from the same student. In order to accomplish this, we randomly choose only one student in each grade from each friendship type cluster so that the estimated effect of peer behavior cannot be identified off of within grade variation. In these estimates, the substantial differences in health behavior across cohorts provide the shock to the health behavior of potential same-grade friends that identifies the effect of friends on health behavior. In practice, we present the average of the parameter estimates resulting from several random draws of one individual in each cluster per cohort.¹⁸

Finally, the rich structure of friendship type clusters, as outlined above, will create singleton clusters of students—those students who have unique or “unusual” friendship preferences for their school. These singleton clusters will, implicitly, not contribute to the identification of the network effects estimates, as there will be no within-cluster variation to exploit. Tables 2A and 3A examine the significance of excluding the variation associated with these observations from our estimates of the effects of friends' health behaviors. While we find some evidence that attrition on this dimension varies with observable attributes, the estimated

¹⁷ In yet another refinement of our cluster approach, in some analyses we also include grade levels-pairs within the clusters, so that 7th and 8th graders are compared to each other (and 9th/10th and 11th/12th) in order to move closer to the thought experiment described in the introduction. All results are robust in terms of magnitude and statistical significance to the use of adjacent cohort comparisons, but the estimates become relatively noisy when we restrict ourselves to one individual per cohort and adjacent clusters, which leads to counterfactuals that are somewhat less informative than those presented in the text.

¹⁸ Standard errors are bootstrapped by repeated sampling of cohorts in schools with replacement, and then for each bootstrap sample we select one observation per cohort repeating this multiple times in order to get an average estimate for each bootstrap sample. We create 1,000 bootstrap samples and estimate the model 30 times per bootstrap sample in order to produce the confidence intervals for our estimate. The repeated averages of estimates from one observation per cohort should not be able to improve the precision of estimates relative to use of the entire sample so when the bootstraps imply smaller standard errors we use the standard errors from the full sample in order to be conservative.

relationship between smoking and drinking status and placement in a single cluster is fairly small (Table 2A). In addition, we repeat the substantive analyses presented below for subsamples excluding observations associated with singleton clusters while holding the fixed effect structure constant (Second row of Table 3A) or varying the fixed effect structure while holding the effect sample constant (Third row of Table 3A), and singleton clusters have little effect on the pattern of estimates observed.

Evidence of Friendship Selection

We can partially test the validity of our approach by examining whether students seem to be sorting into specific friendship patterns within our friendship clusters. Specifically, we test whether a student's own observable attributes correlate with the attributes of their friends within student clusters. Following the logic of Altonji, Elder, and Tabor (2005), if individuals do not sort on observables into friendships within clusters, it is very unlikely that they have sorted based on unobservable characteristics. For example, if we find no evidence of additional correlation between an individual's own parental education and the parental education of their friends after conditioning on the average level of correlation for all students in this cluster, which might include broader educational categories, then it is unlikely that students are sorting based on unobservable characteristics like the parents' involvement with the students' education or the parents' educational and academic expectations since those unobservable characteristics are likely correlated with parental education. Similar diagnostic tests have been used elsewhere (Bayer, Ross and Topa 2008; Bifulco, Fletcher and Ross 2011).

In Table 4A, we present evidence from these diagnostic tests for the effect of friends' smoking and in Table 4B we present results for friends' drinking. Each set of rows examines the correlation between a different "outcome" (individual-level characteristic) and friend's characteristics. Columns add controls from left to right. The first column and row shows the traditional conditional correlation between maternal education and the average level of smoking or drinking among the individual's friends based on OLS estimation.¹⁹ Column 2 controls for school fixed effects, which typically the estimated coefficient by a substantial amount, but the estimated effects are often still sizable and statistically significant at a very high level of confidence. Column 3 controls for school by friendship cluster fixed effects. Column 4 controls

¹⁹ All student demographics used in the creation of friendship clusters are included as regressors in all models unless they are collinear with the fixed effects used in that particular model.

for fixed effects associated with school, friendship cluster, and the individual's race and maternal educational attainment, which is obviously not identified for the case of race, but is identified for maternal years of education. Column 5 presents the results for Column 4 selecting one observation per cohort per cluster and weighting clusters back up to their original size for comparability, except for race where the one observation per cohort sample is estimated using the model in column 3.

In Rows 1 through 6 in Tables 4A and 4B, we examine the conditional correlation between whether white, maternal education, mothers' nativity, whether lives with mother, paternal education and father' nativity and the share friend's smoking or drinking. The OLS and school fixed effect models are sizable and highly significant for all variables except for live with mother, which is statistically significant, but noticeably smaller. However, the inclusion of school by friendship cluster by individual type fixed effects (columns 4 and 5) substantially erodes the school FE estimates. Focusing on the one per cluster estimates, even for the small effects on lives with mother, the estimates are reduced to half the magnitude of the school FE estimates. For the other more sizable estimates, the effects are reduced by between 61 and 97 percent. Of the twelve counterfactuals in column 5, the estimate on percent friends' drinking is just significant at the 5 percent level for nativity of mother and just misses 5% significance for maternal education, but as noted above our method is expected to reduce bias with a finite number of choices and these estimates are substantially smaller than the fixed effect estimates, 68 and 61 percent respectively.

Results

Table 5 presents estimates for adolescent smoking and drinking where same-sex/same-grade friends are used to define the friendship network. In Column 1, the baseline results for smoking (row 1) suggest that increasing the share of friends who smoke by 10 percentage points would increase own-smoking by nearly 3.9 percentage points. In Column 2, we follow some of the previous literature and control for school fixed effects; however this only reduces the coefficient from 0.385 to 0.366 for friends' smoking. In Column 3 we control for school by cluster fixed effects and so control for same friendship-type choices given the same friendship

opportunity set, and we observe a substantially larger decline in the estimated to 0.308.²⁰ The fourth column incorporates individual observables into the cluster definitions. The fifth column presents the results for our preferred specification re-estimating the model in column 4 after restricting the sample to one observation per cohort. The estimates in columns 4 and 5 are very similar to the estimates in Column 3.

All of the estimates based on within friendship cluster by school comparisons fall between 0.308 and 0.320. The lowest estimates are associated with models that contain variation within grade, which is consistent with the possibility that within grade-within cluster variation could bias estimates in the opposite direction from the bias in OLS or school FE models. . Similarly in row 2, school fixed effect estimates of friends' drinking are 0.302, and the estimates after controlling for friendship cluster range between 0.253 and 0.287. Focusing on our preferred estimates using the one observation per cluster per cohort, we see less than a 10% reduction relative to our school fixed effect estimates for both friends' smoking and drinking with our inclusion of individual-friendship type-school fixed effects. Significantly, all reductions due to cluster fixed effects including those based on estimates that include variation within cluster and cohort are less than 20%, which is quite small relative to the declines in estimates across the same model specifications for our balancing tests where the declines are typically on the order of 60 to 90 percent.²¹ As discussed above, as we control for richer cluster definitions, the sample size used to identify the coefficients is reduced due to "singleton clusters". As mentioned earlier and shown in Appendix Tables 3A and 3B, the change in sample composition cannot be the explanation for our persistently strong effects of friends smoking.

The bottom line of these findings is that our estimates suggest only minimal bias from correlated unobservables in school fixed effect estimates of friends' smoking or drinking on a student's own behavior. After controlling for variables like gender, race, maternal education, gender, family structure and nativity, factors that are observed in many educational samples, friendship formation within a grade appears to be relative random at least in terms of unobservables that have a systematic influence on health behaviors. This finding has relevance

²⁰ It is interesting to note that the inclusion of friendship cluster fixed effects produces estimates that are very similar in magnitude to column 2. Rather, it is controlling for friendship choices conditional on the friendship opportunity set (school) that leads to the significant reductions in the estimated effect.

²¹ We also investigate models where smoking or drinking depends upon both friends' smoking and friends' drinking. The estimated effect magnitudes are essentially split between the two friends' behavior variables with the majority of the effect being associated with same behavior: friends' smoking for smoking and friends' drinking for drinking. The pattern over the fixed effects specifications is also similar.

for the growing empirical literature on networks (Calvo-Armengol et al. 2009, Trogdon et al. 2008). Most studies in this area are identified by network fixed effects assuming that individual links within each network are formed randomly. At least after controlling for the role of readily observable demographics in friendship formation, our research is supportive of this assumption for smoking and drinking. However, the assumption of exogenous friendship formation may not be valid for other behaviors or outcomes.

Two Stage Models and an Application to Longitudinal Data

If our estimates are to capture the causal effects of friends' behaviors, an additional assumption is required that a student's own smoking behavior does not directly cause the student to form friendships with students who smoke. We can neither directly test whether this assumption holds or credibly argue that the assumption is reasonable. One reasonable option is to look at models of the on-set of health related behaviors with a longitudinal sample where reverse causality is less of a concern. However, quite frequently, longitudinal datasets are substantially smaller than cross-sectional samples, and the Add Health survey is no exception with the initial in-school survey attempting to interview the full population of each sampled school, while only a fraction of these students are followed over time.

However, the analysis described above generates information on ε_{is} associated with all individuals who are in the same cluster c . Therefore, we can combine this information with the longitudinal data for examining the onset of health behaviors, such as drinking and smoking, which by construction could not have caused the earlier friendship choices. The classic threat to identification in studies of the on-set of health behaviors is that the same set of unobservables that caused the student to select their friends also lead to later smoking, but in our case the information generated on each cluster can provide a control for those correlated unobservables.

Specifically, we estimate the fixed effect associated with each school, student type and friendship pattern cluster, and that fixed effect provides an estimate of the unobservable for each individual within the cluster. However, the fixed effect estimate for each cluster is an average across all cohorts including the cohort from which a student is drawn, which in small samples can lead to an overcorrection from the fixed effects and an understatement of the true estimate. The standard solution to this problem is to calculate the control, in this case the fixed effect, omitting the information that creates this correlation, in this case the information from the same cohort. Guryan, Kroft and Notowidigdo (2009), however, show that the estimated coefficient on

this corrected control may be negatively biased because an individual's own information is negatively correlated with the average of the relevant population (that omits the individual) from which that individual draws peers. In our context of a school fixed effects model, their recommended control for mitigating this bias is simply the average fixed effect estimate across all individuals in the school except for the cohort-cluster to which the subject individual belongs. The resulting estimating equation would be

$$H_{ics} \approx \left(\frac{1}{n_i} \sum_{j \in \Omega_s} H_j \right) \beta_1 + \hat{\rho}_{c-i} + \hat{\rho}_{s-i} + (\mu_{is} - \bar{\mu}_c) + (\rho_c - \hat{\rho}_{c-i} - \hat{\rho}_{s-i}) \quad (11)$$

where $\hat{\rho}_{c-i}$ and $\hat{\rho}_{s-i}$ are the estimated cluster fixed effect omitting the information on the fixed effect from a student's own cohort and the mean (weighted by the number of students) of the cluster fixed effect estimates over all clusters at a school again omitting the student's cohort contribution to their cluster fixed effect estimate.

In order to obtain the most precise estimates for our preferred specification, we collapse the data at the level of the student type, friendship choices, cohort and school in order to obtain estimated based only on cross-cohort information using the entire sample. The parameter estimates and the estimated student type-friendship choice-school FE's are weighted based on the number of students in each student type-friendship choice-cohort-school cluster. We also focus in this analysis on comparisons across all grades in a school, again in order to increase the amount of information used to calculate the FE's. Note that the fixed effects are only identified for friendship clusters in schools containing at least one student in at least two cohorts.

Table 6 presents models from the cross-sectional sample. The first two columns present both the school FE and the student type by friendship choice by school FE model estimates for the individual subsample where one individual is selected randomly per cluster where both are estimated for the one per cohort sample for which fixed effects are estimable. These school FE estimates are very similar to the FE estimates in column 2 of Table 5, and the one observation per cohort per cluster FE's estimates are identical. The next two columns present the student weighted estimates using the collapsed data with school and school by cluster FE's. The school and cluster FE's estimates for these two samples are very similar and continue to indicate very little bias from correlated errors relative to the school FE estimates for smoking or drinking.

In the last two columns, we present the two stage estimates using the individual sample and the predicted cluster FE's. The first of the last two columns includes our prediction of the

cluster FE that includes information from an individual's own cohort. Naturally, the estimate on the fixed effect is large and highly significant because it contains information on the individual's own smoking behavior. The inclusion of this variable reduces the effect of friend's smoking or drinking, representing a classic overcorrection that occurs with fixed effect estimates based on small samples. In the last column, we present the estimates for the fixed effect model that includes the predicted cluster fixed effect that excludes information from the individual's own cohort plus the Guryan et al. (2009) control at the school level. The Guryan et al. control is negative and significant as expected. The estimate on the predicted fixed effect, however, is very near zero and statistically insignificant. As in the single stage fixed effect estimates, there is no evidence of substantial bias from correlated unobservables for friendship formation within schools, and the estimated effect of friends smoking or drinking is quite near to the school FE estimates.

Table 7 presents effect estimates for the onset of smoking or drinking as a function of the behavior of a student's friends at baseline. The first two columns present the OLS and school fixed effect estimates for the full sample of all students surveyed at follow-up who did not smoke or drink, respectively, at baseline. The third column presents the school FE estimates for a sample restricted to those for which a student type by friendship choice by school FE can be estimated (students present in the cluster in that school for at least two cohorts). The fourth column adds the predicted cluster fixed effect using information from the student's own cohort and the fifth column controls for the predicted fixed effect omitting that information and for the Guryan control. The inclusion of school fixed effects erodes the OLS estimates for the friends' effect on the onset of drinking, but has little impact on the effect of friends' smoking on smoking onset. More significantly, the coefficient on the prediction of the cluster FE is effectively zero in the last two columns, and the effect of friends' behavior is relatively unchanged as compared to the school FE presented in column 3. Note that the estimate on the predicted cluster FE are zero even when the own cohort information is included because the smokers and drinkers from the first wave who helped drive these cluster predictions have been deleted from the sample in order to study onset. Similarly, the coefficient estimate on the Guryan control is relatively small in these models when compared to estimates from Table 6.

The central threat to identification in our models of the on-set of smoking and drinking is that the fixed effects may contain considerable noise due to the smaller number of observations

used to estimate each fixed effect. Table 8 presents results based on the size of the school-student type-friendship pattern clusters used to calculate the cluster FE. The average numbers of observations associated with the fixed effect estimates are 2.5 for the subsample of small clusters and 17 for the subsample of big clusters. Obviously, these subsamples are not randomly determined because, for example, clusters involving individuals with mostly own race friends will tend to have a substantially larger number of students than clusters of students who have many friends across racial lines. Regardless, if student friendship formation patterns capture information on unobservables that influence smoking or drinking, then we should find larger, more precisely estimated parameters on the predicted fixed effects when those fixed effects are more accurately measured because they are based on more observations. However, all our results are robust across the subsamples. The estimates on the predicted fixed effects are uniformly small and statistically insignificant, and all our results are robust to the inclusion of the predicted fixed effects. The coefficient on smoking does erode somewhat for the big cluster subsample when the additional controls are included, but we have verified that those changes are driven entirely by the inclusion of the Guryan et al. control, rather than the predicted fixed effect.

Robustness to Number of Friends

One natural concern with our identification strategy is that students only report up to a maximum of five same gender friends and many report only 1 or 2 same gender friends. While students make friendship choices over a wide variety of student attributes that are observable, one still might be concerned that our positive estimates of friendship effects are driven by students who have a very small number of friends and that in those cases the number of choices being made is insufficient to eliminate the bias from student unobservables that influence friendship choice. In Table 9, we divide the sample by the number of friends where the first panel presents estimates for smoking and the second panel presents estimates for drinking. The general pattern of results remains the same with the inclusion of school by friendship cluster fixed effects leading to reductions in estimates by less than 10 percent relative to the school fixed effects for smoking and small increases, less than 5 percent, for drinking. In terms of the magnitudes, the changes in the estimated effect size due to the inclusion of cluster FE's is actually smaller for the subsample of students with 4-5 friends for smoking and drinking in both percentage terms and absolute changes.

In Table 10, we present estimates for the effect of friends' behavior on the on-set of smoking (first panel) and drinking (second panel) for the 1-3 and 4-5 friends subsamples. Since the restricted subsample for 4-5 friends is quite small (between 200 and 300 students), we estimate a joint model for the entire sample interacting dummies for number of friends with the share of friends smoking or drinking, rather than estimate separate models for each subsample.²² In the first two columns, we present the school FE estimates for the individual longitudinal sample. The relationship between friends' behavior and own behavior is robust in both subsamples for both drinking and smoking, and as in cross-sectional estimates of behavioral effects in Table 10 the relationship for on-set is also stronger for the subsample with 4-5 friends. Columns 3 and 4 then present the estimates for the restricted sample. Three of the four significant findings persist in the restricted sample, but the estimate on friend's smoking for the 4-5 friend subsample is very near zero. It is important to note that this zero estimate is for a very small subsample (less than 300 observations), and that the zero estimate does not provide evidence in either support of or against the validity of our identification strategy. Regardless, the inclusion of the predicted cluster FE has very little influence on any of the estimates, reducing the effect in the three positive and significant estimates by less than three percent and changing the zero estimate for friends' smoking for 4-5 friends by only 0.002. Further, we can calculate the predicted cluster FE that does not omit own cohort for the entire longitudinal sample. When we estimate the model including those predicted FE's, the estimated effects of friend's smoking are 0.050 and 0.068 for 1-3 and 4-5 friends, respectively, and those reductions are nearly identical in magnitude across the two subsamples at about 6 percent.

Conclusions

While researchers typically examine peer effects by defining the peer group broadly, this paper focuses attention on actual friends and implements a new research design to study the effects of friend's health behaviors on own health behaviors for adolescents. The main idea is to combine a cross-cohort, within school design with controls for friendship options and friendship choices through the use of "school-student type-friendship pattern" fixed effects. We show that in the Add Health data used in this paper, there is evidence that our design is successful in

²² It is not realistic to estimate 60-70 school fixed effects in samples with only a couple of hundred observations. The pattern and significance arising from models using split samples are qualitatively similar to the estimates in columns 3 and 4, but the magnitudes of the parameter estimates are quite unstable.

narrowing down relevant comparison groups by controlling for the friendship choices and friendship options of adolescents. Our initial estimates also suggest that all results are robust to the restriction of sample to one student per cluster per cohort, which assures that the model is only identified based on comparisons of students across cohorts in the same school.

Further, we use a model of friendship formation to investigate the circumstances under which our identification strategy will provide consistent estimates. We find that our approach can be applied under quite general circumstances. For example, our model allows for a very general non-linear process of friendship selection and allows for correlation between observable attributes and unobservables that affect friendship formation. In addition, we show how to apply the information gained from our analysis to smaller longitudinal samples in order to control for correlated unobservables in models of behavior on-set where the simultaneity between own health behavior and friendship choice is unlikely to be able to explain the estimated effects of friends' behaviors. The key assumptions required to apply this identification strategy are that unobservable determinants of health behavior have a monotonic affect on the patterns of friendship formation and that individuals experience some type of shock in exposure to health behavior of potential friends that does not directly enter own health behavior. This shock assures that some variation remains in friends' health behavior even after asymptotically eliminating variation across individuals in friendship outcomes. In our application, this "treatment" is the variation across cohorts in the exposure to friends' health behavior. Our empirical analysis is very supportive of this assumption in that we find very small variation in the demographic attributes of students across cohorts in the same school, but substantially larger variation in health behavior.

Friends' drinking and smoking appears to have a substantial impact on a student's own smoking and drinking and on the onset of smoking and drinking, and controls for correlated unobservables does little to erode the estimated effect. There continues to be little evidence of bias in the school FE models when we focus on a subsample of students that have the larger numbers of friends or on subsamples of clusters where we observed the largest number of students, which represent circumstances where our analysis should have the greatest potential for identifying bias. Further, counterfactual analyses of race, maternal and paternal education, and maternal and paternal nativity find that at most only weak evidence of an effect of friends'

behavior remains after applying our identification strategy suggesting that our fixed effect strategy eliminates most of the bias from correlated unobservables.

Literature Cited

- Aizer, Anna. 2009. Peer Effects and Human Capital Accumulation: the Externalities of ADD. NBER Working Paper No. 14354.
- Aizer, A, and J Currie. 2004. "Networks or neighborhoods? Correlations in the use of publicly funded maternity care in California." *Journal of Public Economics* 88(12): 2573–85
- Altonji, J G., Elder, TE, CR Tabor. 2005. "Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools." *Journal of Political Economy*, 113: 151-184.
- Angrist, J and K Lang. 2004. "Does school integration generate peer effects? Evidence from Boston's Metco Program." *American Economic Review*, 94(5): 1613-1634.
- Bayer, Patrick, Hjalmarrsson, Randi and David Pozen. 2009. "Building criminal capital behind bars: Peer effects in juvenile corrections," *Quarterly Journal of Economics* 124 (1): 105-147.
- Bayer. P and S. Ross. 2008. "Identifying individual and group effects in the presence of sorting: A neighborhood effects application." *National Bureau of Economic Research Working Paper* #12211
- 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
- Ballester, C., Calvó-Armengol, A. and Y. Zenou. 2006. Who's who in networks. Wanted: the key player. *Econometrica*, 74, 1403-1417
- Bearman, P., Moody J., and K. Stovel. 2004. "Chains of affection: the structure of adolescent romantic and sexual networks." *American Journal of Sociology*, 110 (1): 44-91
- Bertrand, M, E Luttmer, and S Mullainathan. 2000. "Network effects and welfare cultures." *Quarterly Journal of Economics* 115(3):1019–55.
- 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
- Blundell, R and M C. Dias. 2009. "Alternative approaches to evaluation in empirical microeconomics." *Journal of Human Resources*, 44, 565-640.
- Bramouille Y. and B. Rogers. 2009. "Diversity and popularity in social networks. Social Science Research Network, Working Paper, 2009.
- Brock, WA. and SN. Durlauf. 2006. "Identification of binary choice models with social interactions." *Journal of Econometrics*, forthcoming.
- Brock, W A. and S N. Durlauf. 2001. "Discrete choice with social interactions." *Review of Economic Studies*, 68, 235-260.
- Brown, B. B., Dolcini, M. M. and A. Leventhal. 1997. "Transformations in peer relationships at adolescence: implications for health-related behavior.: In: Schulenberg, J., Maggs, J. L. & Hurrelmann, K., eds. *Health Risks and Developmental Transitions During Adolescence*, pp. 161–189. New York: Cambridge University Press.
- Calvó-Armengol, A., Patacchini, E. and Y. Zenou. 2009. "Peer effects and social networks in education." *Review of Economic Studies* 76, 1239-1267
- Case, A and L Katz. 1991. "The company you keep: The effects of family and neighborhood on disadvantaged youths. NBER Working Paper 3705
- Christakis N and J. Fowler. 2007. "The spread of obesity in a large social network over 32 years." *New England Journal of Medicine*, 2007, 357: 370-9.

- Christakis N and J. Fowler. 2008. "The collective dynamics of smoking in a large social network." *New England Journal of Medicine*, 2008, 358: 249-58.
- Clark, A and Y Loheac. 2007. "It wasn't me, it was them!" Social influence in risky behavior by adolescents." *Journal of Health Economics* 26(4): 763-784
- Cohen-Cole, E. and J.M Fletcher. 2008a. "Detecting implausible social network effects in acne, height, and headaches: Longitudinal analysis." *British Medical Journal*, 337: a2533.
- Cohen-Cole, E. and J.M. Fletcher. 2008b. "Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic." *Journal of Health Economics*, 27 (5): 1382-1387.
- Cutler, D. and E. Glaeser. 2007. "Social interactions and smoking." *National Bureau of Economic Research Working Paper* #13477.
- Cutler, David and Edward Glaeser. 1997. Are Ghettos Good or Bad? *Quarterly Journal of Economics*, 112, 827-872.
- Dale, S. and A. Krueger. 2002. "Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables." *Quarterly Journal of Economics*, 117(4):1491-1527.
- 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.
- Evans, William, Wallace Oates, and Robert Schwab. 1992. Measuring Peer Group Effects: A Study of Teenage Behavior. *Journal of Political Economy*, 10, 966-991.
- Fletcher, J.M. 2010. "Social interactions and smoking: Evidence using multiple student cohorts, instrumental variables, and school fixed effects." *Health Economics*. 19(4) 466-484
- Fletcher, J.M. (in press). "Peer Influences on Alcohol Consumption." *Journal of Population Economics*.
- Friesen, J and B Krauth. 2011. "Ethnic Enclaves in the Classroom." *Labour Economics* 18 (5): 656-66.
- Fu, S and S L. Ross. 2010. "Wage premia in employment clusters: How important is worker heterogeneity?" Working Paper
- Gaviria, A and S Raphael. 2001. "School-based peer effects and juvenile behavior." *Review of Economics and Statistics*, 83 (2)
- Glaeser, D L., B Sacerdote, and J A. Scheinkman. 1996. "Crime and social interactions," *Quarterly Journal of Economics*, 111: 507-548.
- Gould, E, V Lavy, and D Paserman. 2009. "Long term classroom peer effects: Evidence from random variation in enrollment of disadvantaged immigrants. *Economic Journal*, 119, 1243-1269.
- 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.
- Halliday, T J. and S. Kwak. 2009. "Weight gain in adolescents and their peers." *Economics and Human Biology*, 7(2), 181-190.
- Halliday, T J. and S. Kwak. 2008. "What is a peer? The role of network definition in the estimation of endogenous peer effects." IZA Discussion Paper #3335.
- Hanushek, E.A., Kain, J.F., Markman, J.M., Rivkin, S.G., 2003. "Does peer ability affect student achievement?" *Journal of Applied Econometrics* 18 (5), 527-544.

- Hoekstra, Mark L. and Scott Carrell. In Press. Externalities in the Classroom: How Children Exposed to Domestic Violence Affect Everyone. *American Economic Journal: Applied Economics*.
- Ioannides, Y M. and L D Loury. 2004. "Job information networks, neighborhood effects, and inequality." *Journal of Economic Literature*, 42, 1056-93.
- Jackson, M.O. Social and Economic Networks. Princeton, NJ: Princeton University Press, 2008.
- Kling, J.R., J.B. Liebman, & L. Katz. (2007). Experimental Analysis of Neighborhood Effects *Econometrica* 75(1): 83-119.
- Lavy, V, D Passerman and A Schlosser. In Press. "Inside the black box of ability peer effects: Evidence from variation in low achievers in the classroom. *Economic Journal*.
- Lavy, VD and A Schlosser. 2011. "Mechanisms and impacts of gender peer effects at school." *American Economic Journal: Applied Economics*. 3(2): 1-33
- Liu X, Lee LF, and J. Kagel. 2006. "Dynamic discrete choice models with lagged social interactions: With an application to a signaling game experiment." Ohio State: Mimeo.
- Lundborg, P. 2006. "Having the wrong friends? Peer effects in adolescent substance use." *Journal of Health Economics*, 25: 214-233
- 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
- Mayer. A and S. Puller. 2008. "The old boy (and girl) network: Social network formation on university campuses." *Journal of Public Economics*, 92 (1-2): 329-347.
- Patacchini, E. and Y. Zenou. In Press. "Juvenile delinquency and conformism." *Journal of Law, Economics and Organization*.
- Powell, L, J Taurus, and H Ross. 2005. "The Importance of peer effects, cigarette prices, and tobacco control policies for youth smoking behavior." *Journal of Health Economics*, 24: 950-968.
- Renna, F, I Grafova, and N Thakur. 2008. "The effects of friends on adolescent body weight." *Economics & Human Biology* 6, (3): 377-387.
- Ross, S L. 2011. "Social interactions within cities: Neighborhood environments and peer relationships." In *Handbook of Urban Economics and Planning* (Eds. N. Brooks, K. Donaghy, G. Knapp). Oxford University Press.
- Sacerdote. B. 2001. "Peer effects with random assignment: Results for Dartmouth roommates." *Quarterly Journal of Economics*, 116 92): 681-704.
- Trogdon, J, J Nonnemaker, J Pais. 2008. "Peer effects in adolescent overweight." *Journal of Health Economics*, 27(5): 1388-1399
- Topa, G. 2001. "Social interactions, local spillovers, and unemployment." *Review of Economic Studies*, 68: 261-295.
- Weinberg, B. 2008. "Social interactions with endogenous associations." Ohio State University Working Paper

Table 1
Descriptive Statistics

Variable	Mean	Std Dev	Min	Max
Smoke	0.34	0.47	0	1
Drink	0.54	0.50	0	1
Get Drunk	0.29	0.45	0	1
Exercise	2.29	1.19	0	4
Any Exercise	0.95	0.22	0	1
Age	14.87	1.69	10	19
Male	0.45	0.50	0	1
White	0.62	0.49	0	1
Hispanic	0.13	0.33	0	1
Black	0.16	0.37	0	1
Asian	0.06	0.23	0	1
Live with Mom	0.94	0.25	0	1
Maternal Years of Education	13.49	2.33	0	18
Maternal Caring Scale	4.79	0.58	1	5
Native Born	0.92	0.26	0	1
Grade = 7	0.14	0.35	0	1
Grade = 8	0.14	0.35	0	1
Grade = 9	0.21	0.41	0	1
Grade = 10	0.19	0.39	0	1
Grade = 11	0.16	0.37	0	1
Grade = 12	0.15	0.36	0	1
Missing	0.34	0.47	0	1
Number of Nominations	2.78	1.30	1	5
Proportion White	0.61	0.43	0	1
Proportion Black	0.17	0.35	0	1
Proportion Hispanic	0.13	0.28	0	1
Proportion Asian	0.06	0.20	0	1
Proportion Other Race	0.04	0.14	0	1
Proportion Mom Less High School	0.14	0.25	0	1
Proportion Mom Some College	0.18	0.25	0	1
Proportion of Mom College Grad	0.35	0.33	0	1
Proportion Native	0.93	0.22	0	1
Proportion Live with Mom	0.94	0.18	0	1

Notes: This table presents the mean, standard deviation, minimum and maximum for the approximately 54,000 individuals sampled in wave 1 who nominated at least one friend of the same gender who could be confirmed to attend the same grade in the same school and reported whether they smoke or drank.

Table 2
Distribution of Health Behaviors in Friendship Networks

% Smoke	Freq.	Percent	Cum.	% Drink	Freq.	Percent	Cum.
Same Sex Friends							
0.00	22,994	42.51	42.51	0.00	12,509	23.18	23.18
0.10				0.10			
0.20	1,534	2.84	45.34	0.20	931	1.73	24.91
0.30	7,270	13.44	58.78	0.30	5,542	10.27	35.18
0.40	1,154	2.13	60.91	0.40	1,064	1.97	37.15
0.50	7,146	13.21	74.12	0.50	7,713	14.3	51.45
0.60	770	1.42	75.55	0.60	1,135	2.1	53.55
0.70	2651	4.9	80.45	0.70	3,774	6.99	60.55
0.80	1,748	3.23	83.68	0.80	3,440	6.38	66.92
0.90				0.90			
1.00	8,830	16.32	100	1.00	17,847	33.08	100
Total	54,097	100		Total	53,955	100	

Notes: The results identify the number and frequency of students who have friends with an average incidence of health behavior falling into a particular bin. The bins run from the number listed to any number less than the number in the next row, but since individuals at most have five same gender friends the top and bottom rows capture the number of people with friends where either none smoke or drink or all smoke or drink. Similarly, the rows associated with 0.20-0.29 and 0.80-89 both captures individuals with five friends where four exhibit one behavior and the other does not, but 0.20-0.29 also captures individuals with four friends where one smokes. Sample sizes are based on the total number of individuals with at least one same grade/same gender friend who reported their drinking or smoking behavior.

Table 3
Variation in Friendship Options

Peer Variable	R-squared
% Maternal College Graduate	92.5%
% Black	97.2%
% Hispanic	97.4%
% White	
% Asian	93.8%
% Native Born	96.1%
Mean Maternal Caring Scale	55.1%
% Smoke Cigarettes	76.5%
% Drink Alcohol	80.9%

Notes: The results reported indicate the R-squared from a regression of the grade-level measure of peer characteristics or peer health behaviors on a complete set of school-level and grade-level dummy variables. The sample is the approximately 65,000 students who were surveyed in wave 1.

Table 4A
Balancing Tests of Friendship Sorting: Friends' Smoking

Outcome Specification	White OLS				
Fixed Effects	None	School	School-Cluster	School-Cluster-X	School-Cluster-X One Per Cohort
% Smoke	0.171*** (0.027)	0.077*** (0.015)	0.007 (0.009)	NA	0.002 (0.009)
Observations	53,564	53,564	53,564		43,197
R-squared	0.017	0.365	0.808		0.952
Outcome Specification	Maternal Education OLS				
Fixed Effects	None	School	School-Cluster	School-Cluster-X	School-Cluster-X One Per Cohort
% Smoke	-0.292*** (0.075)	-0.324*** (0.048)	-0.227*** (0.060)	-0.088* (0.050)	-0.077 (0.050)
Observations	53,564	53,564	53,564	53,564	43,236
R-squared	0.002	0.112	0.602	0.883	0.951
Outcome Specification	Nativity of Mother OLS				
Fixed Effects	None	School	School-Cluster	School-Cluster-X	School-Cluster-X One Per Cohort
% Smoke	0.056*** (0.018)	0.028*** (0.006)	0.010* (0.005)	0.004 (0.006)	0.002 (0.006)
Observations	52,322	52,322	52,322	52,322	42,197
R-squared	0.010	0.163	0.698	0.822	0.880

Table 4A
Balancing Tests of Friendship Sorting: Friends' Smoking

Outcome Specification	Live w/ Mom OLS	Live w/ Mom OLS	Live w/ Mom OLS	Live w/ Mom OLS	Live w/ Mom OLS School-Cluster-X One Per Cohort
Fixed Effects	None	School	School-Cluster	School-Cluster-X	
% Smoke	-0.010** (0.004)	-0.012*** (0.003)	-0.012** (0.006)	-0.008* (0.005)	-0.006 (0.005)
Observations	52,761	52,761	52,761	52,761	42,474
R-squared	0.001	0.011	0.537	0.890	0.927
Outcome Specification	Paternal Education OLS	Paternal Education OLS	Paternal Education OLS	Paternal Education OLS	Paternal Education OLS School-Cluster-X One Per Cohort
Fixed Effects	None	School	School-Cluster	School-Cluster-X	
% Smoke	-0.299*** (0.061)	-0.269*** (0.045)	-0.216*** (0.067)	-0.124 (0.082)	-0.110 (0.082)
Observations	40,517	40,517	40,517	40,517	32,450
R-squared	0.004	0.087	0.619	0.775	0.853
Outcome Specification	Nativity of Father OLS	Nativity of Father OLS	Nativity of Father OLS	Nativity of Father OLS	Nativity of Father OLS School-Cluster-X One Per Cohort
Fixed Effects	None	School	School-Cluster	School-Cluster-X	
% Smoke	0.082*** (0.023)	0.028*** (0.007)	0.015* (0.008)	0.005 (0.008)	0.005 (0.008)
Observations	40,760	40,760	40,760	40,760	32,662
R-squared	0.008	0.312	0.779	0.882	0.920

Table 4B
Balancing Tests of Friendship Sorting: Friends' Drinking

Outcome Specification	White OLS				
Fixed Effects	None	School	School-Cluster	School-Cluster-X	School-Cluster-X One Per Cohort
% Drink	0.071*** (0.024)	0.027** (0.010)	-0.001 (0.008)	NA	0.001 (0.008)
Observations	53,445	53,445	53,445		43,108
R-squared	0.003	0.361	0.809		0.949
Outcome Specification	Maternal Education OLS				
Fixed Effects	None	School	School-Cluster	School-Cluster-X	School-Cluster-X One Per Cohort
% Drink	-0.239*** (0.057)	-0.276*** (0.034)	-0.176*** (0.048)	-0.074 (0.045)	-0.088* (0.045)
Observations	53,445	53,445	53,445	53,445	43,095
R-squared	0.002	0.111	0.603	0.882	0.910
Outcome Specification	Nativity of Mother OLS				
Fixed Effects	None	School	School-Cluster	School-Cluster-X	School-Cluster-X One Per Cohort
% Drink	0.067*** (0.017)	0.049*** (0.010)	0.023*** (0.008)	0.020** (0.009)	0.019** (0.009)
Observations	52,212	52,212	52,212	52,212	42,085
R-squared	0.012	0.166	0.700	0.822	0.868

Table 4B
Balancing Tests of Friendship Sorting: Friends' Drinking

Outcome Specification	Live w/ Mom OLS	Live w/ Mom OLS	Live w/ Mom OLS	Live w/ Mom OLS	Live w/ Mom OLS School-Cluster-X One Per Cohort
Fixed Effects	None	School	School-Cluster	School-Cluster-X	
% Drink	-0.009** (0.004)	-0.009** (0.004)	-0.009 (0.006)	-0.007 (0.006)	-0.005 (0.006)
Observations	52,647	52,647	52,647	52,647	42,397
R-squared	0.001	0.011	0.538	0.890	0.926
Outcome Specification	Paternal Education OLS	Paternal Education OLS	Paternal Education OLS	Paternal Education OLS	Paternal Education OLS School-Cluster-X One Per Cohort
Fixed Effects	None	School	School-Cluster	School-Cluster-X	
% Drink	-0.259*** (0.063)	-0.275*** (0.043)	-0.159** (0.074)	-0.091 (0.091)	-0.070 (0.091)
Observations	40,412	40,412	40,412	40,412	32,417
R-squared	0.003	0.088	0.619	0.775	0.853
Outcome Specification	Nativity of Father OLS	Nativity of Father OLS	Nativity of Father OLS	Nativity of Father OLS	Nativity of Father OLS School-Cluster-X One Per Cohort
Fixed Effects	None	School	School-Cluster	School-Cluster-X	
% Drink	0.066*** (0.019)	0.036*** (0.007)	0.014* (0.007)	0.007 (0.009)	0.005 (0.009)
Observations	40,665	40,665	40,665	40,665	32,659
R-squared	0.006	0.313	0.779	0.882	0.918

Notes: Each set of rows and each column in Tables 4A and 4B display coefficients from separate regressions. Additional controls include age, gender, race and ethnicity dummy variables, maternal education, whether live with mother, whether native born and indicators for whether an observation has data missing on a given variable. Cluster and X refer to cells associated with friendship patterns and student demographics, respectively. Paternal and maternal attributes are based on student reports where education is years of schooling and nativity is whether not born in the U.S. All regressions control for grade-level fixed effects, and standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1.

Table 5
Friendship Network Effects

Outcome Specification	Smoke OLS				
Fixed Effects	None	School	School-Cluster	School-Cluster-X	School-Cluster-X One Per Cohort
% Smoke	0.385*** (0.010)	0.366*** (0.011)	0.308*** (0.019)	0.301*** (0.023)	0.320 (0.023)
Observations	49,845	49,845	49,845	49,845	41723
R-squared	0.140	0.147	0.582	0.703	0.803
Outcome Specification	Drink OLS				
Fixed Effects	None	School	School-Cluster	School-Cluster-X	School-Cluster-X One Per Cohort
% Drink	0.329*** (0.011)	0.302*** (0.011)	0.253*** (0.019)	0.256*** (0.022)	0.287 (0.022)
Observations	49,656	49,656	49,656	49,656	41572
R-squared	0.153	0.163	0.609	0.726	0.821

Notes: Each column displays coefficients from separate regressions. Cluster and X refer to cells associated with friendship patterns and student demographics, respectively. All regressions control for grade-level fixed effects and indicators for observations where control variables are missing. Standard errors (shown in parentheses) are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1.

Table 6
Two Stage Estimates of Friendship Effects on Smoking and Drinking

Outcome Specification	Smoke School FE	Smoke School-Cluster-X One Per Cohort	Smoke School FE	Smoke School-Cluster-X	Smoke School FE, Cluster FE Est. w/ own Cohort	Smoke School FE, Cluster FE Est. w/out own Cohort
Sample	Individual	Individual	Collapsed	Collapsed	Individual	Individual
% Smoke	0.340*** (0.014)	0.320*** (0.023)	0.363*** (0.017)	0.336*** (0.027)	0.318*** (0.010)	0.328*** (0.011)
Predicted FE w/ or w/out own cohort					1.037*** (0.003)	-0.002 (0.009)
Guryan et. al. Control						-171.477*** (16.991)
Observations	11,373	11,373	11,373	11,373	17,500	17,500
R-squared	0.147	0.539	0.192	0.553	0.377	0.392
Outcome Specification	Drink School FE	Drink School-Cluster-X One Per Cohort	Drink School FE	Drink School-Cluster-X	Drink School FE, Cluster FE Est. w/ own Cohort	Drink School FE, Cluster FE Est. w/out own Cohort
Sample	Individual	Individual	Collapsed	Collapsed	Individual	Individual
% Drink	0.289*** (0.014)	0.287*** (0.022)	0.320*** (0.016)	0.321*** (0.023)	0.287*** (0.010)	0.288*** (0.012)
Predicted FE w/ or w/out own cohort					1.037*** (0.004)	0.005 (0.011)
Guryan et. al. Control						-179.230*** (15.565)
Observations	11,340	11,340	11,340	11,340	17,435	17,435
R-squared	0.154	0.550	0.209	0.568	0.385	0.403

Notes: Each column and panel displays coefficients from separate regressions. The individual sample is estimated using student level data for which a school by type by friendship pattern fixed effect is identified and the first two columns are based on the one student per cohort per fixed effect sample, while the collapsed sample is based on one observation for each grade-school-cluster-X cell and weighted by the number of students in that cell. The predicted fixed effect regressor in Column 5 is the estimated school-cluster-X fixed effect from the model shown in column 4. The predicted fixed effect in Column 6 is the weighted mean over all grades of the grade-school-cluster-X residual from the model in column 4 omitting the fixed effect contribution from a student's own grade. The Guryan control is the same weighted mean over all grade-cluster-X cells in a school again omitting the contribution from a students' own grade-cluster-X cell. All regressions control for grade-level fixed effects, standard demographic controls shown above and indicators for observations where variables are missing. Standard errors (shown in parentheses) are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1.

Table 7
Longitudinal Analysis of the onset of Smoking and Drinking

Outcome Specification	Smoke OLS	Smoke School FE	Smoke School FE	Smoke School FE, Cluster FE Est. w/ own Cohort	Smoke School FE, Cluster FE Est. w/out own Cohort
Sample	Full Sample	Full Sample	Restricted	Restricted	Restricted
% Smoke	0.063*** (0.011)	0.061*** (0.012)	0.053*** (0.016)	0.051*** (0.017)	0.045*** (0.017)
Predicted FE w/ or w/out own cohort				-0.018 (0.035)	0.002 (0.021)
Guryan et. al. Control					11.743** (5.392)
Observations	6,137	6,137	1,895	1,895	1,895
R-squared	0.019	0.041	0.093	0.093	0.095
Outcome Specification	Drink School FE	Drink School FE	Drink School FE	Drink School FE, Cluster FE Est. w/ own Cohort	Drink School FE, Cluster FE Est. w/out own Cohort
Sample	Full Sample	Full Sample	Restricted	Restricted	Restricted
% Drink	0.120*** (0.015)	0.112*** (0.016)	0.110*** (0.027)	0.107*** (0.028)	0.114*** (0.028)
Predicted FE w/ or w/out own cohort				-0.044 (0.056)	-0.028 (0.028)
Guryan et. al. Control					-10.836 (11.298)
Observations	4,310	4,310	1,228	1,228	1,228
R-squared	0.034	0.074	0.119	0.120	0.121

Notes: Each column and panel displays coefficients from separate regressions. The restricted sample is estimated only using students who belong to school-cluster-X cells where there are students in that cell from at least two grades (fixed effects identified). The predicted fixed effect in Column 4 is the estimated school-cluster-X fixed effect from the model shown in column 4 of Table 6. The predicted fixed effect in Column 5 is the weighted mean over all grades of the grade-school-cluster-X residual from the same model omitting the fixed effect contribution from a student's own grade. The Guryan control is the same weighted mean over all grade-cluster-X cells in a school again omitting the contribution from a students' own grade-cluster-X cell. All regressions control for grade-level fixed effects, standard demographic controls shown above and indicators for observations where variables are missing. Standard errors (shown in parentheses) are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1.

Table 8
 Longitudinal Analysis of the onset of Smoking and Drinking
 Stratified by Cluster Size

Outcome Specification	Smoke School FE	Smoke School FE	Smoke School FE, Cluster FE Est. w/out own Cohort	Smoke School FE, Cluster FE Est. w/out own Cohort
Sample	Restricted Small Cluster	Restricted Large Cluster	Restricted Small Cluster	Restricted Large Cluster
% Smoke	0.055* (0.029)	0.052** (0.021)	0.047 (0.031)	0.041** (0.021)
Predicted FE w/out own cohort			-0.025 (0.029)	0.054 (0.036)
Guryan et. al. Control			9.439 (11.739)	18.294** (8.465)
Observations	833	1,062	833	1,062
R-squared	0.190	0.125	0.191	0.133
Outcome Specification	Drink School FE	Drink School FE	Drink School FE, Cluster w/out own Cohort	Drink School FE, Cluster w/out own Cohort
Sample	Restricted Small Cluster	Restricted Large Cluster	Restricted Small Cluster	Restricted Large Cluster
% Drink	0.151*** (0.048)	0.079** (0.037)	0.151*** (0.047)	0.082** (0.038)
Predicted FE w/out own cohort			-0.015 (0.044)	-0.034 (0.070)
Guryan et. al. Control			-1.374 (18.174)	-24.185* (14.256)
Observations	571	657	571	657
R-squared	0.238	0.175	0.238	0.179

Notes: Each column and panel displays coefficients from separate regressions. The restricted sample is estimated only using students who belong to school-cluster-X cells where there are students in that cell from at least two grades. Small and large clusters refer to subsamples with the number of observations in school-cluster-X cell above or below the median. The predicted fixed effect is the weighted mean over all grades of the grade-school-cluster-X residual from the model in column 4 of Table 7 omitting the fixed effect contribution from a student's own grade. The Guryan control is the same weighted mean over all grade-cluster-X cells in a school again omitting the contribution from a students' own grade-cluster-X cell. All regressions control for grade-level fixed effects, standard demographic controls shown above and indicators for observations where variables are missing. Standard errors (shown in parentheses) are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1.

Table 9
Friendship Effects by Number of Friends

Outcome Specification	Smoke School FE	Smoke School FE	Smoke School-Cluster-X FE	Smoke School-Cluster-X FE
Sample	Collapsed Sample 1-3 Friends	Collapsed Sample 4-5 Friends	Collapsed Sample 1-3 Friends	Collapsed Sample 4-5 Friends
% Smoke	0.326*** (0.017)	0.599*** (0.031)	0.307*** (0.027)	0.589*** (0.058)
Observations	9,201	2,154	9,201	2,154
R-squared	0.187	0.284	0.546	0.604
Outcome Specification	Drink School FE	Drink School FE	Drink School-Cluster-X FE	Drink School-Cluster-X FE
Sample	Collapsed Sample 1-3 Friends	Collapsed Sample 4-5 Friends	Collapsed Sample 1-3 Friends	Collapsed Sample 4-5 Friends
%Drink	0.290*** (0.016)	0.539*** (0.033)	0.297*** (0.025)	0.543*** (0.063)
Observations	9,173	2,149	9,173	2,149
R-squared	0.201	0.313	0.559	0.621

Notes: Each column and panel displays coefficients from separate regressions. Each column represents a subsample of individuals with either 1-3 or 4-5 same grade/same gender friends. Cluster and X refer to cells associated with friendship patterns and student demographics, respectively. The collapsed sample is based on one observation for each grade-school-cluster-X cell and weighted by the number of students in that cell. All regressions control for grade-level fixed effects, standard demographic controls shown above and indicators for observations where variables are missing. Standard errors (shown in parentheses) are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1.

Table 10
Friendship Effects on On-set by Number of Friends

Outcome Specification	Smoke School FE	Smoke School FE	Smoke School FE	Smoke School FE	Smoke School FE- Cluster FE Est. w/out own Cohort	Smoke School- Cluster FE Est. w/out own Cohort
Sample	Individual 1-3 Friends	Individual 4-5 Friends	Restricted 1-3 Friends	Restricted 4-5 Friends	Restricted 1-3 Friends	Restricted 4-5 Friends
% Smoke	0.053*** (0.015)	0.073*** (0.017)	0.067*** (0.020)	0.001 (0.029)	0.066*** (0.021)	-0.001 (0.029)
Observations		6,137		1,895		1,895
R-squared		0.042		0.094		0.094
Outcome Specification	Drink School FE	Drink School FE	Drink School FE	Drink School FE	Drink School FE- Cluster FE Est. w/out own Cohort	Drink School- Cluster FE Est. w/out own Cohort
Sample	Individual 1-3 Friends	Individual 4-5 Friends	Restricted 1-3 Friends	Restricted 4-5 Friends	Restricted 1-3 Friends	Restricted 4-5 Friends
% Drink	0.088*** (0.016)	0.151*** (0.029)	0.083*** (0.028)	0.220*** (0.071)	0.081*** (0.029)	0.217*** (0.073)
Observations		4,310		1,228		1,228
R-squared		0.075		0.124		0.124

Notes: Each pair of columns in each panel displays coefficients from separate regressions. The restricted sample is estimated only using students who belong to school-cluster-X cells where there are students in that cell from at least two grades. Each column represents a subsample of individuals with either 1-3 or 4-5 same grade/same gender friends. Columns 5 and 6 include controls for a predicted fixed effect based on a weighted mean over all grades of the grade-school-cluster-X residual from the model in column 4 of Table 7 omitting the fixed effect contribution from a student's own grade and a Guryan-style control that is the same weighted mean over all grade-cluster-X cells in a school again omitting the contribution from a student's own grade-cluster-X cell. All regressions control for grade-level fixed effects, standard demographic controls shown above and indicators for observations where variables are missing. Standard errors (shown in parentheses) are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1.