Letter

Cheuk Yin Ho*

The Relationship from Friendship Links to Educational Achievement

DOI 10.1515/bejeap-2015-0267

Published online April 8, 2016

Abstract: Using the pseudo-panel structure of friendship links, this paper employs a control function approach to estimate the impact of the number of friends on one's own educational achievement in linear models. Results show that having one additional reciprocal or out-degree friend increases about 2% of a standard deviation in cognitive test scores.

Keywords: friendships, test scores, control functions

JEL Classification: I20, I21, A14

1 Introduction

In the literature of peer effects, there are some recent papers investigating the effects of the number of friends on one's own outcomes, such as labor productivity (Conti et al. 2013; Fletcher 2014), educational achievement (Mihaly 2009; Navy and Sand 2014), and health (Ho 2016). To distinguish the causal effect of the number of friends from unobserved heterogeneity, Mihaly (2009) uses an instrumental variables approach by excluding school composition variables from the achievement equation to obtain identification; Conti et al. (2013) and Ho (2016) use a structural approach to jointly model friendship formation and outcomes to account for selection; Fletcher (2014) use within-siblings estimation to wipe out family-specific unobservables across siblings; Navy and Sand (2014) exploit a random assignment of students to classrooms to examine the impact of classroom-friends on test scores. Except for Mihaly (2009), the other studies find a positive effect of the number of friends on labor productivity or educational achievement or health.

This paper uses a new empirical strategy to estimate the impact of the number of friends (sum of friendship links) on one's own educational achievement. Exploiting the pseudo-panel structure of friendship links, this paper

*Corresponding author: Cheuk Yin Ho, Division of Business and Management, BNU-HKBU United International College, 28 Jinfeng Road Tangjiawan, Zhuhai, China,
E-mail: efcyho@gmail.com

adopts a control function approach to deal with unobserved individual heterogeneity. In the first step, I exploit the pseudo-panel structure of friendship data, in which individuals make multiple binary friendship choices at a given period of time, to estimate individual effects in a linear dyad model. In the second step, the estimated individual effects are included as a control function in the education production, which controls for individual unobservables, so the effect of the number of friends on educational achievement is identified.

Unlike Conti et al. (2013) and Ho (2016), the identification of this paper does not rely on the non-linearity of a friendship model. In addition, this paper does not assume any parametric distribution of individual unobservables. Instead, the source of identification comes from the pseudo-panel structure of friendship links that is exploited to estimate individual effects. The estimated individual effects are not perfectly collinear in observable variables in the linear education production function, because the pseudo-panel (school size) varies across individuals. Excluding homophily variables in the dyad model from the education production further sharpens the identification.

Estimating models with data from the National Longitudinal Study of Adolescent Health (Add Health), this paper finds two main results. First, including individual effects estimated from the dyad model significantly increases the effects of the number of friends on test scores. This suggests that this paper's empirical strategy helps to reduce a downward bias in the OLS estimate. Second, results show that having one additional reciprocal or out-degree friend increases about 2% of a standard deviation in cognitive test scores.

The remainder of this paper is organized as follows. Section 2 describes the data. Section 3 outlines the empirical strategy. Section 4 discusses the empirical results, and Section 5 concludes.

2 Data

The data used in this paper come from the National Longitudinal Study of Adolescent Health (Add Health). Individuals were asked to nominate their friends (up to five males and five females) in the 1994–1995 in-school survey. Friends not included in school rosters cannot be identified and thus are excluded from the study, which implies that an individual's friends are also his or her schoolmates. In the empirical analysis, I define out-degree friends as the individual's schoolmates who were nominated by the individual as friends; I define in-degree friends as the individual's schoolmates who nominate the individual as friends; I also define reciprocal friends as the individual's schoolmates who nominate and are nominated by the individual as friends. In the

in-home survey, individuals were asked to fill out a long-form questionnaire in which their individual, parental, and neighborhood particulars were collected. In addition, they were asked to take the Add Health Picture Vocabulary Test (AHPVT), which assessed their verbal ability and scholastic aptitude. The summary statistics of variables are given in Table 1.

Table 1: Summary statistics.

	Mean	Std. Dev.	Min	Max	Observation
<i>Friendship variable</i>					
Number of out-degree friends	4.01	3.00	0	10	13,466
Number of in-degree friends	4.14	3.65	0	33	13,466
Number of reciprocal friends	1.55	1.67	0	9	13,466
<i>Outcome variable</i>					
AHPVT	98.7	15.1	9	137	12,814
<i>Individual covariate</i>					
Male	0.49	0.50	0	1	13,566
Age	16.2	1.68	12.0	21.1	13,566
Minority	0.41	0.49	0	1	13,566
<i>Parental covariate</i>					
Parental college education	0.39	0.49	0	1	11,512
Live with mother	0.92	0.27	0	1	13,232
Live with father	0.74	0.44	0	1	13,169
Number of siblings	0.86	1.03	0	6	12,673
Mother professional	0.27	0.45	0	1	12,658
Father professional	0.17	0.38	0	1	12,595
Living condition	3.36	0.84	1	4	13,406
<i>Neighborhood covariate</i>					
Population in block/100	17.6	15.0	0	348	13,449
Fraction minority in block	0.29	0.33	0	1	13,444
Fraction urban in block	0.59	0.48	0	1	13,444
Fraction aged 25 + W/O HS diploma in block	0.29	0.16	0	0.93	13,436
Fraction aged 25 + W college degree in block	0.22	0.14	0	0.94	13,436
Unemployment rate in block	0.08	0.06	0	0.5	13,199
Median household income in block/1000	30.6	13.8	5.0	130	13,021

3 Empirical Methodology

The education production function is structured as follows:

$$Y_i = \beta F_i + X_i' \Theta + \rho \gamma_i + \eta_i$$

[1]

where Y_i is the educational achievement of individual i . The variable of interest is the number of friends F_i , where β captures the associated effect on the achievement. X_i is a vector of exogenous covariates and η_i is the i.i.d. error term. F_i is endogenous because it is correlated with the unobserved variable γ_i , i. e., $\rho \neq 0$.¹ Thus, estimating eq. [1] by ordinary least-squares (OLS) gives a biased estimate of β .

The key to identifying β is to control for the unobservable γ_i . The traditional control function approach (Heckman and Robb 1985) is to run a regression of F_i on exogenous variables X_i and IV_i and then include the residual as a regressor to control for γ_i in eq. [1]. Successful identification requires IV_i to be correlated with F_i but excluded from eq. [1]; otherwise, the residual is perfectly collinear in X_i and F_i , which fails the rank condition. The novelty of this paper is to exploit the pseudo-panel structure of friendship links to help achieve identification.

To understand the identification, a dyadic model of friendship links is set up as follows:

$$F_{ij} = Z'_{ij}\Pi + \gamma_i + \alpha_j + \epsilon_{ij} \quad [2]$$

$$F_i = \sum_{j=1}^{J_s} F_{ij} \quad [3]$$

where $F_{ij} = 1$ if i forms a friendship link with j and zero otherwise. Friendship links are either directed (i. e., F_{ij} might not equal F_{ji}) or undirected (i. e., $F_{ij} = F_{ji}$). Links are directed for out-degree friends and in-degree friends, but undirected for reciprocal friends. Self-ties are ruled out (i. e., $F_{ii} = 0$). Z_{ij} are dyad observables that characterize homophily; I parameterize Z_{ij} as $|Z_i - Z_j|$, which measures absolute demographic differences in race, gender, and school grade between i and j , namely, $|Minority_i - Minority_j|$, $|Male_i - Male_j|$, and $|Grade_i - Grade_j|$, respectively.² γ_i captures individual characteristics that determine friendship decisions, such as race, gender, parental background, cognitive ability, and personality traits, etc. α_j captures peer effect: individual j is attractive to others and then highly valued to be friends. ϵ_{ij} is a random shock of match quality that is independently and identically distributed across dyads. γ_i and α_j are treated as parameters to be estimated, so the joint distribution of unobserved heterogeneity and observed variables of homophily is left unrestricted. In particular,

¹ Reverse causality is not an issue because individuals made their friends a long time ago before taking the AHPVT, so the AHPVT performance cannot causally affect the number of friends.

² This kind of parameterization is standard in the social network literature, see Goldsmith-Pinkham and Imbens (2013), Graham (2015), and Hsieh and Lee (2016)

observable covariates, X_i , are absorbed by γ_i . I normalize $\sum \gamma_i = 0$ and $\sum \alpha_j = 0$. I further assume that $\eta_i \perp (\gamma_i, F_i, X_i)$, $\gamma_i \perp (Z_{ij}, X_i)$, $\epsilon_{ij} \perp (Z_{ij}, X_i, \gamma_i, \alpha_j)$, and $\epsilon_{ij} \sim i.i.d. (0, \sigma^2)$.

From eq. [3], an individual's number of school friends is the sum of his or her friendship links: $F_i = \sum_{j=1}^{J_s} F_{ij}$, where J_s is the number of schoolmates in school s . It is clear that F_i is endogenous in eq. [1] because γ_i jointly affects friendship decisions and achievement. Owing to multiple observations of binary friend choices of each individual, γ_i can be identified in eq. [2]. In a linear model, it follows that $\hat{\gamma}_i = \frac{\sum_{j=1}^{J_s} F_{ij}}{J_s} - \frac{\sum_{j=1}^{J_s} Z'_{ij} \hat{\Pi}}{J_s} = \frac{F_i}{J_s} - \tilde{Z}'_i \hat{\Pi}$, where $\tilde{Z}_i = \frac{\sum_{j=1}^{J_s} Z_{ij}}{J_s}$. The following are the details of the two-step estimation strategy:

Step 1: The linear probability model is used to estimate $\hat{\gamma}_i$ in eq. [2].³

Step 2: $\hat{\gamma}_i$ is included as a control function in eq. [1].

In a linear model, the estimated individual effects, $\hat{\gamma}_i = \frac{F_i}{J_s} - \tilde{Z}'_i \hat{\Pi}$, are not perfectly collinear in X_i and F_i in eq. [1] because of two identifying sources. First, J_s is not constant: school sizes are heterogeneous. Second, $\tilde{Z}_i$ could be excluded from the achievement equation. $\tilde{Z}_i$ includes average absolute differences between an individual and his or her schoolmates in race, gender, and school grade. Conditional on school fixed effects, which subsume demographic characteristics of schoolmates, $\tilde{Z}_i$ might not directly affect an individual's achievement. Nevertheless, exclusion restrictions are not required for identification. Thus, I construct two different control functions. Control function 1 is constructed as $\hat{\gamma}_i^1 = F_i/J_s$ without $\tilde{Z}'_i \hat{\Pi}$. Control function 2 is constructed as $\hat{\gamma}_i^2 = \frac{F_i}{J_s} - \tilde{Z}'_i \hat{\Pi}$, where homophily variables averaged at the individual level $\tilde{Z}_i$ are excluded from eq. [1].

4 Results

Table 2 reports estimation results. To ease interpretations, test scores are standardized with a mean of zero and a standard deviation of one. To take into account of estimation errors of the coefficients in the linear probability model, bootstrapped standard errors are used in the model using control function 2. The specification in column 1 includes school fixed effects, grade fixed effects,

³ For friendship data, the panel is long in the sense of both large N and J , i. e., the average school size is 659 students, so the fixed effects estimator is consistent and asymptotically normal as $N, J \rightarrow \infty$.

Table 2: Estimation results.

	Dependent variable: AHPVT		
	(1)	(2)	(3)
Out-degree friends	0.018*** (0.004)	0.027*** (0.006)	0.027*** (0.005)
Control function 1		−0.051** (0.020)	
Control function 2			−0.037*** (0.011)
School fixed effects	✓	✓	✓
Grade fixed effects	✓	✓	✓
Individual covariates	✓	✓	✓
Parental covariates	✓	✓	✓
Neighborhood covariates	✓	✓	✓
Missing-variables dummies	✓	✓	✓
Observation	12,652	12,652	12,652

Note: Robust standard errors clustered by schools in parentheses in columns (1) and (2); Bootstrapped standard errors in parentheses in column (3).

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

individual, parental, and neighborhood covariates, as well as missing-variable dummies. It shows that having one out-degree friend increases one’s own test score by 1.8% of a standard deviation.

To deal with unobserved individual heterogeneity, the specification in column 2 adds control function 1 to the model. It shows that the coefficient of interest significantly changes from 0.018 to 0.027. The coefficient of control function 1 is -0.051 and statistically significant. It means that a one-standard-deviation increase in the individual unobservable lowers one’s own test score by 5.1% of a standard deviation. The result is interpreted as follows: First, there is a tradeoff between leisure and study hours – individuals who spend more time in social activities like making friends have less time and motivation in their studies. Hence, omitting such factors in the education production generates a downward bias in the OLS estimate. This result is in line with Conti et al. (2013) who find that the individual unobservable of friendship choices has a negative impact on labor productivity. Second, measurement errors of the number of friends shrink the OLS estimate toward zero, but the control function approach helps to correct the downward attenuation bias. For a robustness check, control function 2 is used in column 3. The estimated coefficient of control function

becomes smaller, but the estimated effect of the number of out-degree friends is unchanged.

Table 3 shows results from a model in which three types of number of friends are jointly included. Control functions 1 and 2 are conditioned in columns 1 and 2, respectively. In column 1, the result shows that having one additional reciprocal or out-degree friend increases about 2% of a standard deviation in cognitive test scores. However, having in-degree friends has no effect. The results are robust to using control function 2 in column 2.

Table 3: Estimation results.

	Dependent variable: AHPVT	
	(1)	(2)
Out-degree friends	0.022*** (0.006)	0.022*** (0.006)
In-degree friends	−0.004 (0.004)	−0.004 (0.004)
Reciprocal friends	0.019** (0.008)	0.019** (0.008)
Control function 1	✓	
Control function 2		✓
School fixed effects	✓	✓
Grade fixed effects	✓	✓
Individual covariates	✓	✓
Parental covariates	✓	✓
Neighborhood covariates	✓	✓
Missing-variables dummies	✓	✓
Observation	12,652	12,652

Note: Robust standard errors clustered by schools in parentheses in column (1); Bootstrapped standard errors in parentheses in column (2).
*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

In the empirical analysis, individuals were restricted from nominating more than five male and five female friends. This is not a critical issue because only 8 % of individuals nominated five male friends, 12.5 % of individuals nominated five female friends, and 2.5 % of individuals nominated ten friends. For a robustness check, friendship dummies are defined such that the values are equal to one when individuals have one or more friends and zero otherwise. These alternative friendship variables are not constrained by the upper bound of friendship

Table 4: Robustness check.

	AHPVT
Out-degree friend dummy	0.084*** (0.032)
In-degree friend dummy	−0.006 (0.030)
Reciprocal friend dummy	0.055** (0.024)

Note: $N=12,652$. All models include school fixed effects, grade fixed effects, individual covariates, parental covariates, neighborhood covariates, missing variables dummies, and control function 1. Robust standard errors clustered by schools in parentheses.
*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

nominations. Table 4 shows that the qualitative results do not change – out-degree and reciprocal friends matter for individual achievement. However, the magnitude of the estimates is larger using friendship dummies.

5 Conclusion

This paper uses a control function approach to estimate the impact of the number of friends (sum of friendship links) on one’s own educational achievement. Results show that the empirical approach helps to reduce a downward bias in the OLS estimate. The main finding is that having one additional reciprocal or out-degree friend increases about 2% of a standard deviation in cognitive test scores.

Acknowledgment: I am grateful to an editor and an anonymous referee for helpful comments. This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (<http://www.cpc.unc.edu/addhealth>). No direct support was received from grant P01-HD31921 for this analysis.

Appendix

The appendix shows heterogeneous effects of the number of male and female friends on males and females, respectively. Table 5 shows that having more out-degree male friends boosts test scores of both males and females. In contrast, either male or female in-degree friends have no impact on one’s own test score.

Table 5: Heterogeneous effects.

	Dependent variable: AHPVT	
	(1) Male	(2) Female
Out-degree male friends	0.028*** (0.008)	0.022** (0.010)
Out-degree female friends	0.025*** (0.009)	0.014 (0.012)
In-degree male friends	−0.004 (0.008)	−0.007 (0.006)
In-degree female friends	−0.003 (0.007)	−0.006 (0.009)
Reciprocal male friends	0.005 (0.015)	0.059*** (0.018)
Reciprocal female friends	0.015 (0.017)	0.019 (0.014)

Note: $N = 12,652$. All models include school fixed effects, grade fixed effects, individual covariates, parental covariates, neighborhood covariates, missing variables dummies, and control function 1. Robust standard errors clustered by schools in parentheses.

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

References

Conti, G., A. Galeotti, G. Muller, and S. Pudney. 2013. “Popularity.” *Journal of Human Resources* 48 (4):1072–94.

Fletcher, J. 2014. “Friends or Family? Revisiting the Effects of High School Popularity on Adult Earnings.” *Applied Economics* 46 (20):2408–17.

Goldsmith-Pinkham, P., and G. Imbens. 2013. “Social Networks and the Identification of Peer Effects.” *Journal of Business and Economic Statistics* 31 (3):253–64.

Graham, B. 2015. “Methods of Identification in Social Networks.” *Annual Review of Economics* 7:465–85.

Heckman, J., and R. Robb. 1985. “Alternative Methods for Evaluating the Impact of Interventions: An Overview.” *Journal of Econometrics* 30:239–67.

- Ho, C. Y. 2016. "Better Health with More Friends: The Role of Social Capital in Producing Health." *Health Economics* 25 (1):91–100.
- Hsieh, C. -S., and L. F. Lee. 2016. "A Social Interactions Model with Endogenous Friendship Formation and Selectivity." *Journal of Applied Econometrics* 31 (2):301–19.
- Mihaly, K. 2009. "Do More Friends Mean Better Grades? Student Popularity and Academic Achievement." Rand Working Paper 678.
- Navy, V., and E. Sand. 2014. "The Effect of Social Networks on Student's Achievement and Non-Cognitive Behavioral Outcomes: Evidence from Conditional Random Assignment of Friends in School." Working Paper.