NAME YOUR FRIENDS, BUT ONLY FIVE? THE IMPORTANCE OF CENSORING IN PEER EFFECTS ESTIMATES USING SOCIAL NETWORK DATA

Alan Griffith*

Empirical peer effects research often employs censored peer data. Individuals may only list a fixed number of links, implying mismeasured peer variables. I first document that censoring is widespread in network data. I then introduce an estimator and characterize its inconsistency analytically; an assumption on the ordering of peers implies that censoring causes attenuated peer effects estimates. Next, I demonstrate the effect of censoring in two datasets, showing that estimates with censored data under-estimate peer influence. I discuss interpretation of estimates, propose a corrective method, and give implications for the design of network surveys.

1. INTRODUCTION

In studying the impact of peers in networks, researchers often collect data on individuals' links, or peers. To this end, surveys typically ask individuals to identify other individuals with whom they are linked, often their "friends." A common feature of this data, however, is the presence of censoring, whereby the survey design limits the number of links that individuals may list. A common question allows individuals to name up to five friends; if an individual has, for example, eight, the final three friends are censored and thus unobserved in the data, leading to mismeasurement of the relevant peer group. To this point, there has been little effort to detail the implications of this mismeasurement on estimation, or to develop methods for correction. This paper fill this gap.

First, I document that censored network data is quite prevalent, as shown by the proportion of individuals who name the maximum allowed number of friends across a wide number of surveys. Next, I derive analytic expressions for inconsistency in a commonly-used reduced-form linear-in-means estimator of peer effects. To derive closed-form expressions, I define the property of *order irrelevance*, which requires that the distribution of peer characteristics does not vary across nomination ranks. This implies, importantly, that unobserved links' characteristics follow the same distribution as observed links'. When order irrelevance holds, peer influence parameters are attenuated, while the direction of inconsistency for parameters on one's own characteristics depends on the sign of these effects and the relationship between one's own and one's peers' characteristics. In the pres-

Version of February 5, 2021. Department of Economics, University of Washington. E-mail: alangrif@uw.edu

*This paper has benefited immensely from extensive discussions with and comments from Eric Auerbach, Vincent Boucher, Rachel Heath, Tyler McCormick, Jeff Smith, and Rebecca Thornton, as well as feedback from the 2018 Midwest International Economic Development Conference, the 5th Annual NSF Network Science in Economics Conference, and seminars at the University of Michigan, University of Washington, University of California-Berkeley ARE, and Ohio State University. All errors remain my own.

ence of homophily, whereby individuals' characteristics are positively correlated with those of their friends, the inconsistency on parameters of one's own characteristics is in the opposite direction to the inconsistency of peer parameters. Magnitude of inconsistency depends upon the difference between true and censored average inverse number of links.

I next demonstrate censoring-induced inconsistency in two real datasets. In both applications, I *sub-censor* the data, whereby for a given censoring rule $k = 1, ..., 5$, I ignore all link nominations greater than k. Results are clear: the more links that are censored, the more attenuated are peer effects parameters; that is, peer effects parameters trend away from zero as more links are observed. As implied by order irrelevance and homophily, parameters on own characteristics trend toward zero as more links are observed. Further, consistent with the analytic results, when the relevant variable is randomly assigned, the estimated effect of one's own covariate is unaffected by censoring.

A major implication of the analytic and empirical results is that, on the whole, peer effects parameters using censored network data are *under*-estimated in magnitude. Further, even in the absence of explicitly censored data, noisily-measured peer networks may lead to attenuation of peer effects estimates. This has important implications for how researchers should interpret peer effects estimates.

I go farther, however, by proposing a corrective strategy that relies upon a restricted version of the model. These corrective methods rely crucially on order irrelevance, which allows for consistent estimation of features of the bias-correction term from the observed, censored data. Paired with supplemental data on the true degree distribution, this allows for estimation of bias corrected parameters. I demonstrate bias correction in both of the empirical applications, showing how the (artificially) sub-censored estimates can be corrected to match the least-censored estimates available from the data. I then discuss implications for data collection, suggesting that sufficient data for bias correction may be collected either via two-step edgewise sampling (Conley and Udry, 2012) or Aggregate Relational Data (Breza et al., 2020; McCormick and Zheng, 2015).

The results here relate to a series of papers that investigate the impact of missing data on peer effects estimates. First, in classroom settings, where peer groups are defined as all other students within each individual's classroom, Ammermueller and Pischke (2009) and Sojourner (2013) demonstrate that standard methods may be inconsistent when many students' covariates are unknown, and they suggest corrective methods. Similar to the analytic results here, Ammermueller

and Pischke (2009) derive a formula that relates the inconsistency of a widely-used IV estimator to the percentage of missing peers. In contrast, Boucher et al. (2014) show that a common fixed effect estimator is consistent in a setting with group-based interaction, as the missing peer data is the same for all agents in any group.

In contrast, in network settings, the identity—and in many cases, the number—of peers is unknown. Accordingly, this paper is most closely related to a series of papers that analyze the importance of network mismeasurement on various estimands.[1] Hardy et al. (2019) analyze the estimation of binary treatment effects when networks are measured with error. The correction method proposed here is related to solutions proposed by Breza et al. (2020) and Boucher and Houndetoungan (2019) that use supplemental information about the network (or the distribution of networks) to recover peer effects parameters in the absence of complete network data. In sharp contrast to the bulk of these papers—and to the method proposed by Chandrasekhar and Lewis (2011)—the methods developed here have the advantage of avoiding the requirement of specification and estimation of a structural model of network formation, which can be quite complex and computationally-intensive (See, e.g. Badev, 2017; Griffith, 2019; Mele, 2017).

2. CENSORING IN NETWORK SURVEYS

A widespread practice in collecting network data is to ask individual respondents to name their network links. This practice dates at least to AddHealth, which elicits network data with the following prompt:

> List your closest male friends. List your best male friend first, then your next best friend, and so on. Girls may include boys who are friends and boyfriends (Harris, 2009).

In AddHealth, each student in the sample is given this prompt as well as a prompt to list female friends. Crucially, students are allowed to list only up to five male and five female friends.

Prompts of this type are widespread, especially in surveys collected as part of field projects in developing countries. For instance, Banerjee et al. (2012) allow respondents to list up to either five or eight links along a number of dimensions. Oster and Thornton (2012), Cai, de Janvry and Sadoulet (2015), and Kandpal and Baylis (2013) allow respondents to name up to three, five, and

[1] A series of papers also analyzes the importance of missing network data on features of the network itself. Hoff et al. (2013) and Fosdick and Hoff (2015) develop methods of inference for features of the network itself given ranked-choice but censored data as dealt with here. Thirkettle (2019) develops methods for identification and estimation of network features via a structural model of network formation.

five friends, respectively.

Since it represents an upper bound on the number of censored links, we can get a rough idea of the extent of censoring by looking at the number of respondents who list the maximum allowable number of links. At one end of the spectrum, Banerjee et al. (2012) show that less than 0.1% of their survey respondents name the maximum number of links (either 5 or 8) along any of the many dimensions they survey.[2] However, this appears to be an exception. In the AddHealth Wave 1 in-school data, 66.1% of female respondents and 56.1% of male respondents nominate the maximum number of same-gender friends, while 37.4% of female respondents nad 49.9% of male respondents nominate the maximum number of opposite-gender friends. In the data from Cai, de Janvry and Sadoulet (2015), used in Section 4, the mjority of respondents list the maximum five links, with an average of 4.9. In Oster and Thornton (2012), 68% of sampled girls report the maximum allowed three friends.

While censoring is quite common in network surveys, it is not universal. Two separate approaches seek to collect uncensored network data. First, some have addressed the issue of censoring by collecting ordered connection data but with no upper bound on the number of links. In this vein, Ngatia (2015) and Comola and Prina (2018) allow respondents to list relationships without limit. Second, rather than asking individuals to list contacts, other researchers have prompted individuals to provide their relationship to identified other individuals. Delavallade, Griffith and Thornton (2016) and Tjernström (2017) conduct pairwise network censuses, whereby each individual is queried about their relationship with each other individual within a defined group. Pairwise censuses may be impractical when group sizes are large, which may necessitate asking individuals about random others, as was done in Conley and Udry (2012).

3. CHARACTERIZING INCONSISTENCY ANALYTICALLY

3.1. *Data-Generating Process*

3.1.1. *Networks*

Agents affect each other through a network of links. The only restrictions I impose are that network links are binary, and they need not be symmetric. That is, $l_{ij} \in \{0, 1\}$ gives the existence

[2]In this setting, censoring is unlikely to cause much of an issue, although partial sampling at the node (household) level is still an issue in that dataset (Chandrasekhar and Lewis, 2011).

of a link between agents i and j.[3] By convention, elements along the diagonal are all zeros ($l_{ii} = 0$). From this, we construct a row-normalized adjacency matrix $\mathbf{G}$, where each element $g_{ij} = \frac{l_{ij}}{\sum_{k \neq i} l_{ik}}$ and thus the elements in each row of $\mathbf{G}$ sum to 1.[4] In order that $\mathbf{G}$ be well-defined, assume that every agent has at least one link.

3.1.2. *Outcomes Conditional on the Network*

Define $y_i \in \mathbb{R}$ as some outcome for individual i, while $x_i \in \mathbb{R}^m$ is a vector of characteristics for the same individual. Next, define $\bar{y}_i = \sum_{j \neq i} g_{ij} y_j \in \mathbb{R}$ as the mean of agent i's links' outcomes and $\bar{x}_i = \sum_{j \neq i} g_{ij} x_{js} \in \mathbb{R}^m$ as the mean of agent i's links' characteristics. Outcomes are determined through the linear-in-means process in Equation (1) (see, e.g. Manski, 1993).

$$y_i = \beta_0 + \beta_1 \bar{y}_i + x_i' \beta_2 + \bar{x}_i' \beta_3 + \epsilon_i \tag{1}$$

Peer effects enter through $\bar{y}_i$, the mean of individual i's links' outcome, and $\bar{x}_i$, the mean of those same peers' exogenous characteristics.[5] Written in matrix notation, Equation (1) becomes

$$\mathbf{y} = \beta_0 \iota + \beta_1 \mathbf{G}\mathbf{y} + \mathbf{x}\beta_2 + \mathbf{G}\mathbf{x}\beta_3 + \epsilon \tag{2}$$

where $\mathbf{y}$ is the vector of all outcomes y_i, $\mathbf{x}$ is a matrix of covariates such that row i corresponds to x_i', and ι is an $N \times 1$ vector in which each entry is 1.

ASSUMPTION 1 $\mathbb{E}[\epsilon_i | \mathbf{x}, \mathbf{G}] = 0$

Assumption 1 provides the primary exogeneity assumption. This assumes independence of each agent i's unobserved ϵ_i from i's own observed x_i and the observed characteristics of others. Further, and crucially, it assumes network exogeneity: unobservables that play a part in forming networks $\mathbf{G}$ are not correlated with unobservables in the outcome equation ϵ.[6] This exogeneity assumption

[3] There is a limited number of papers, mostly theoretical, that allow for continuous links (see, e.g., Bloch and Dutta, 2009; Baumann, 2017; Griffith, 2019).

[4] For notational convenience, I have suppressed the group coefficient s. Asymptotic results rely upon the assumption that we observe a large number S of independent groups (such as schools). This in turn implies that $\mathbf{G}$ is block diagonal, since links do not exist across groups.

[5] The discussion follows if we also add in group-level fixed effects γ_s to control for "correlated effects." Group-level fixed effects are included in the empirical analyses in Section 4.

[6] It also bears noting that Assumption 1 provides no restriction on the relationship between $\mathbf{G}$ and $\mathbf{x}$, and the discussion of the direction of inconsistency under homophily explicitly assumes that the network $\mathbf{G}$ may be related

may be quite strong in many cases. However, even with this very strong assumption on the data generating process, estimates of parameters of the model may still be inconsistent due to censoring.

I further make boundedness assumptions on the parameter space, covariates, and unobserved variables, in Assumption 2.

ASSUMPTION 2 *The following boundedness conditions hold*

[1] $(\beta_0, \beta_1, \beta_2, \beta_3) \in \mathbf{B}$*, where* $\mathbf{B}$ *is a compact set in* $\mathbb{R}_{2+2m}$ *such that* $|\beta_1| < 1$*.*

[2] For all i, s*,* $(x_i, \epsilon_i) \in \mathbf{X} \times \mathbf{E}$ *a compact set in* $\mathbb{R}_{m+1}$

An immediate implication of Assumption 2 is that $\mathbf{Gx}$, $\mathbf{y}$, and $\mathbf{Gy}$ are bounded.

3.1.3. *What is Observed*

First, I define what is observed without censoring. Define d_i^{true} as agent i's *degree*, the total number of links held by agent i. Links have an ordering of some type denoted by $(v) = 1, ..., d_i^{true}$.[7] Define $\bar{x}_{i(1)} \in \mathbb{R}^m$ as the characteristics of i's first listed friend, $\bar{x}_{i(2)}$ of the second friend, etc.[8] Therefore, $\bar{x}_i = \frac{1}{d_i^{true}} \sum_{v=1}^{d_i^{true}} \bar{x}_{i(v)}$ and $\bar{y}_i = \frac{1}{d_i^{true}} \sum_{v=1}^{d_i^{true}} \bar{y}_{i(v)}$.

When there is censoring of agent i's network, we observe fewer links than d_i^{true}. For a given censoring rule k, define d_i^k as that agent's degree with network data censored at k. So, $d_i^1 \leq d_i^2 \leq ... \leq d_i^{true}$.[9] Whenever $d_i^k < d_i^{true}$, then agent i's network is censored for a censoring rule k. Next, define $\mathbf{H_k}$ as the row-normalized adjacency matrix induced by censoring rule k. From this, we construct the censored matrix of average friends' characteristics $\mathbf{H_k x}$. In this notation, row i of $\mathbf{H_k x}$ is given by $\frac{1}{d_i^k} \sum_{v=1}^{d_i^k} \bar{x}'_{i(v)}$.

The analysis here deals with consistency of estimators. In all cases, I assume that the number of clusters—whether they be villages, classrooms, schools, or other groupings—is growing without bound. That is, I assume that we observe infinitely many clusters of finitely-many actors each of

to covariates $\mathbf{x}$.

[7]If links are unordered, as is common in many datasets (see, e.g. Cai, de Janvry and Sadoulet, 2015), then assume a random ordering and the discussion follows.

[8]For clarity, this means that if i's first listed friend is agent j, then $\bar{x}_{i(1)} = x_j$.

[9]Note that, for a given k, d_i^k may be greater than k, for multiple reasons. If the peer group mean is constructed using an "OR" definition—whereby a link between i and j exists if either lists the other as a friend—then agent i listing k links but some other agents j listing a link to i that i does not list will lead to $d_i^k > k$. Further, the censoring rule k may refer to the maximum number of links along multiple dimensions, as in AddHealth which allows listing up to five male and five female friends. If agent i lists 5 male and 3 female friends, then $d_i^5 = 8$, even before accounting for others listing i as a friend.

which is independent, in contrast to the literature looking at asymptotics dealing with a single, infinitely growing set of connected individuals (see Graham, 2015; Leung, 2019).

3.2. *The Reduced-Form Estimator without Censoring*

The object of interest is a reduced-form estimator of the linear-in-means model, which I describe here. Assume for simplicity that y_i, x_i, and $\bar{x}_i$ have been demeaned (and thus $\beta_0 = 0$) either in the entire population or by group s, where the latter case corresponds to a groupwise fixed-effects transformation. The additional assumption $|\beta_1| < 1$ allows for rearrangement of Equation (2) as

$$\mathbf{y} = (I - \beta_1 \mathbf{G})^{-1}(\beta_2 \mathbf{x} + \beta_3 \mathbf{G}\mathbf{x} + \epsilon) \tag{3}$$

Definition 1 gives the RF estimator $\hat{\alpha}_{RF}$.

DEFINITION 1 *The RF Estimator* $\hat{\alpha}_{RF} = \left([\mathbf{x}, \mathbf{G}\mathbf{x}]'[\mathbf{x}, \mathbf{G}\mathbf{x}]\right)^{-1} [\mathbf{x}, \mathbf{G}\mathbf{x}]'\mathbf{y}$

That is, $\hat{\alpha}_{RF}$ gives the coefficients of a regression of the outcome y_i on x_i and $\bar{x}_i$. Next, in Definition 2, I define the reduced-form parameter α as the probability limit of $\hat{\alpha}_{RF}$.

DEFINITION 2 *The reduced-form peer effects parameter* $\alpha = \operatorname{plim} \hat{\alpha}_{RF}$

To simplify notation, define the following $m \times m$ matrices: $\mathbb{E}[\mathbf{x}'\mathbf{x}] = \mathbf{E_{xx}}$, $\mathbb{E}[\mathbf{x}'\mathbf{G}\mathbf{x}] = \mathbf{E_{xG}}$, $\mathbb{E}[\mathbf{x}'\mathbf{G}'\mathbf{G}\mathbf{x}] = \mathbf{E_{GG}}$. With these defined, Proposition 1 gives α, the probability limit of the RF estimator in the absence of censoring.

PROPOSITION 1 *Given Assumptions 1-2,*

$$\alpha = \begin{bmatrix} \beta_2 \\ \beta_3 + \beta_1\beta_2 \end{bmatrix} + \begin{bmatrix} \mathbf{E_{xx}} & \mathbf{E_{xG}}' \\ \mathbf{E_{xG}} & \mathbf{E_{GG}} \end{bmatrix}^{-1} \sum_{r=1}^{\infty} \beta_1^r \begin{bmatrix} \mathbb{E}[\mathbf{x}'\mathbf{G}^{r+1}\mathbf{x}] \\ \mathbb{E}[\mathbf{x}'\mathbf{G}'\mathbf{G}^{r+1}\mathbf{x}] \end{bmatrix} (\beta_3 + \beta_1\beta_2)$$

See Appendix A for the proof. In general, α does not admit a simple closed form, as it depends on two features of the data-generating process. First, it depends on the structural parameters $\beta = (\beta_1, \beta_2, \beta_3)$. Second, α depends on the relationships among exogenous $\mathbf{x}$, peer group mean $\mathbf{G}\mathbf{x}$,

and higher-order means $\mathbf{G^r x}$ (friends of friends, friends of friends of friends, etc.).[10] In the common case of homophily (on variable x_i) in link formation, the mean of agents' peers' characteristics $\bar{x}_i$ is (positively) correlated with own characteristics x_i. Therefore, $\mathbb{E}[x_i \bar{x}_i] > 0$ (in matrix notation, $\mathbb{E}[\mathbf{x'Gx}] > 0$).

The result in Proposition 1 nests some well-known special cases. First, when there is no endogenous peer effect ($\beta_1 = 0$), $(\alpha_1, \alpha_2) = (\beta_2, \beta_3)$. Second, when $\mathbf{G} = \mathbf{G}^2$, as in the classroom model with an inclusive (not leave-one-out) mean, $(\alpha_1, \alpha_2) = (\beta_2, \frac{\beta_3 + \beta_1 \beta_2}{1 - \beta_1})$, a standard result given in, e.g., Manski (1993) and Carrell, Sacerdote and West (2013).

3.3. *With Censoring*

Censoring arises when $d_i^k < d_i^{true}$ for at least some agent i. That is, with data censored at k, we do not observe $\mathbf{G}$ and thus cannot directly estimate $\hat{\alpha}_{RF}$. Rather, we estimate $\hat{\alpha}_{RF}^{cens,k}$, which is calculated using $\mathbf{H_k}$ rather than $\mathbf{G}$. Its definition is given in Definition 3.

DEFINITION 3 *The Censored RF Estimator* $\hat{\alpha}_{RF}^{cens,k} = ([\mathbf{x}, \mathbf{H_k x}]'[\mathbf{x}, \mathbf{H_k x}])^{-1}[\mathbf{x}, \mathbf{H_k x}]'\mathbf{y}$

To simplify notation, define the following matrices, noting that all are $m \times m$: $\mathbf{E_{xH}} = \mathbb{E}[\mathbf{x'H_k x}]$, $\mathbf{E_{HH}} = \mathbb{E}[\mathbf{x'H_k' H_k x}]$, and $\mathbf{E_{HG}} = \mathbb{E}[\mathbf{x'H_k' G x}]$. Next, define the following $2m \times 2m$ matrices: $\mathbf{C_{GG}} = \begin{bmatrix} \mathbf{E_{xx}} & \mathbf{E_{xG}} \\ \mathbf{E_{xG}'} & \mathbf{E_{GG}} \end{bmatrix}$, $\mathbf{C_{HH}} = \begin{bmatrix} \mathbf{E_{xx}} & \mathbf{E_{xH}} \\ \mathbf{E_{xH}'} & \mathbf{E_{HH}} \end{bmatrix}$, and $\mathbf{C_{HG}} = \begin{bmatrix} \mathbf{E_{xx}} & \mathbf{E_{xG}} \\ \mathbf{E_{xH}'} & \mathbf{E_{HG}} \end{bmatrix}$. Note that $\mathbf{C_{HH}}$ and $\mathbf{C_{HH}}$ are dependent on k, which is suppressed for notational convenience.

PROPOSITION 2 *Given Assumptions 1-2, for any k,*

$$\begin{aligned} \operatorname{plim} \hat{\alpha}_{RF}^{cens,k} = \mathbf{C_{HH}}^{-1}\mathbf{C_{HG}} \, (\alpha \\ + \sum_{r=1}^{\infty} \beta_1^r \left(\mathbf{C_{HG}}^{-1} \mathbb{E}[\mathbf{x}, \mathbf{H_k x}]' \mathbf{G}^{r+1} \mathbf{x}] - \mathbf{C_{GG}}^{-1} \mathbb{E}[\mathbf{x}, \mathbf{G x}]' \mathbf{G}^{r+1} \mathbf{x}] \right) (\beta_1 \beta_2 + \beta_3) \end{aligned}$$

The probability limit of $\hat{\alpha}_{RF}^{cens,k}$ is given in Proposition 2.The proof is in Appendix A. In general,

[10] This definition of α provides a generalization of the standard sufficiency result for the existence of peer effects. From the structural model, peer effects exist if *either* $\beta_1 \neq 0$ or $\beta_3 \neq 0$. From the probability limit above, we see that a sufficient condition for this is that $\alpha_2 \neq 0$. Note that this is a one-way implication, and there are combinations of parameters such that $\alpha_2 = 0$ yet peer effects still are present.

it is clear that the difference between plim $\hat{\alpha}_{RF}^{cens,k}$ and α depends crucially on the difference between true network $\mathbf{G}$ and censored network $\mathbf{H_k}$, and that as $\mathbf{H_k} \to \mathbf{G}$, $\hat{\alpha}_{RF}^{cens,k} \to \alpha$. That is, as there is less censoring, the probability limit of the censored estimator approaches the true parameter. Further, two additional special cases bear noting. Corollary 1 provides that, when networks are independent of the covariate vector $\mathbf{x}$, then the estimator $\hat{\alpha}_1^{cens,k}$ is consistent even in the presence of censoring.

COROLLARY 1 *Given Assumptions 1-2, for any k, if* $\mathbb{E}[\mathbf{x'H_kx}] = \mathbb{E}[\mathbf{x'Gx}] = 0$, *then* plim $\hat{\alpha}_1^{cens,k} = \alpha_1$.

This condition will hold when, for example, the covariate is a randomly-assigned treatment indicator. In this common case, the censored estimate of the parameter on *own* treatment remains consistent despite mismeasured networks due to censoring.

Finally, when there is no endogenous peer effect ($\beta_1 = 0$), then the inconsistency of $\hat{\alpha}^{cens,k}$ has a much simpler form, as given in Corollary 2. This condition is crucially important in deriving simple characterizations of inconsistency as well as correction strategies.

COROLLARY 2 *Given Assumption 1-2, for any k, if* $\beta_1 = 0$, *then* plim $\hat{\alpha}^{cens,k} = \mathbf{C_{HH}}^{-1}\mathbf{C_{HG}}\alpha$.

3.4. *Inconsistency under Order Irrelevance*

In general, the formula for inconsistency in Proposition 2 does not have a simple closed form. However, an assumption on the relationship between friends' x and the order they are named gives restrictions that aid in pinning down the direction of inconsistency due to censoring, as well as providing conditions for correction. The key assumption is a property called *order irrelevance*, and its definition is given in Assumption 3.

ASSUMPTION 3 *(Order Irrelevance) The following hold for all* $v, w \neq v$, *for all* $d_i^{true}, d_i^{cens,k}$:

[1] $\mathbb{E}[x_i \bar{x}'_{i(v)} | d_i^{true}, d_i^{cens,k}] = \mathbf{A}$

[2] $\mathbb{E}[\bar{x}_{i(v)} \bar{x}'_{i(v)} | d_i^{true}, d_i^{cens,k}] = \mathbf{B}$

[3] $\mathbb{E}[\bar{x}_{i(v)} \bar{x}'_{i(w)} | d_i^{true}, d_i^{cens,k}] = \mathbf{C}$

Assumption 3 provides three restrictions. First, the covariance of an individual's own characteristics x and the individuals' links' characteristics is constant across orderings. Thus, for example, if

x_i is an indicator for individual i being male, then his first listed friend is equally likely to also be male as his last listed friend, and equally likely to be male as censored friends. Second, the variance of each friend's characteristics, given by $\mathbb{E}[\bar{x}_{i(v)}\bar{x}'_{i(v)}]$, is constant across friendship orderings. Third, the covariance between any two friends' characteristics is also independent of ordering. That is, for example, the covariance between the characteristics of first and second friends is the same as the covariance between the first and seventh, etc.

Intuitively, order irrelevance means that the distribution of peers' characteristics are the same across different nomination orderings. This, in turn, implies that the distribution of characteristics of friends that we do not observe due to censoring is the same as the distribution of friends' characteristics that we do. The payoff of Assumption 3 is that it allows for clear predictions on the direction of inconsistency, and it motivates correction methods discussed in Section 5.

The result with order irrelevance is given in Proposition 3. Note that this is much simpler than the general result in Corollary 2. Crucially, note that the magnitude of inconsistency depends on $\left(\mathbb{E}[\frac{1}{d^{cens,k}}] - \mathbb{E}[\frac{1}{d^{true}}]\right)$, the difference between average inverse degree in the censored data and true average inverse degree.

PROPOSITION 3 *Given Assumptions 1 - 3, if* $\beta_1 = 0$*, then*

$$\operatorname{plim} \hat{\alpha}_{RF}^{cens,k} = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} + \left(\mathbb{E}[\frac{1}{d^{cens,k}}] - \mathbb{E}[\frac{1}{d^{true}}]\right) \begin{bmatrix} \mathbf{E_{xx}}^{-1}\mathbf{A} \\ -\mathbf{I} \end{bmatrix} \left(\mathbf{E_{HH}} - \mathbf{A'}\mathbf{E_{xx}}^{-1}\mathbf{A}\right)^{-1} (\mathbf{B} - \mathbf{C})\, \beta_3$$

A special but important case is when the variable $\mathbf{x}$ is randomly assigned (at the individual level). Random assignment implies that there is no correlation between each agent's x_i and others' in i's network.[11] Therefore, $\mathbf{A} = 0$. Random assignment further implies no correlation *among* each agent's links. That is, if i is linked to both j and k, then $\mathbb{E}[x_{js}x'_{ks}] = 0$, and $\mathbf{C} = 0$. This insight leads to Corollary 3, which follows directly from Proposition 3.

[11] I note that, with small groups, random assignment without replacement may lead to "exclusion bias," which arises due to a mechanical (negative) correlation between one's own treatment status and one's peers' treatment (Caeyers and Fafchamps, 2019). Assuming $\mathbf{A} = 0$ essentially assumes no exclusion bias.

COROLLARY 3 *Given Assumptions 1-3, if* $\beta_1 = 0$ *and if* $\mathbf{A} = \mathbf{C} = 0$*, then for any* k*,*

$$\operatorname{plim} \hat{\alpha}_{RF}^{cens,k} = \begin{bmatrix} \beta_2 \\ \frac{\mathbb{E}[\frac{1}{d^{true}}]}{\mathbb{E}[\frac{1}{d^{cens,k}}]}\beta_3 \end{bmatrix}$$

Since for any censoring rule k, $d_i^{cens,k} \leq d_i^{true}$, it follows that $\mathbb{E}[\frac{1}{d^{true}}] \leq \mathbb{E}[\frac{1}{d^{cens,k}}]$. Therefore, if the conditions of Corollary 3 hold, the peer effect is necessarily attenuated: its probability limit is in the same direction but (weakly) smaller in magnitude than the true value.

COROLLARY 4 *Given Assumptions 1 - 3, if* $\beta_1 = 0$ *and* $m = 1$

$$\operatorname{plim} \hat{\alpha}_{RF}^{cens,k} = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} + (\mathbb{E}[\frac{1}{d^{cens,k}}] - \mathbb{E}[\frac{1}{d^{true}}]) \begin{bmatrix} \mathbf{A} \\ -\sigma_x^2 \end{bmatrix} \frac{\mathbf{B} - \mathbf{C}}{\sigma_x^2 \mathbf{E_{HH}} - \mathbf{A}^2}\beta_3$$

where $\sigma_x^2 = \mathbb{E}[x_i^2]$.

A further special case, given in Corollary 4, arises when $m = 1$, in which case the covariate is a scalar. This provides clear implications for the direction of inconsistency. First, $(\mathbb{E}[\frac{1}{d^{cens,k}}] - \mathbb{E}[\frac{1}{d^{true}}]) \geq 0$ always. Second, due to the Cauchy-Schwartz Inequality, it must be true that $\mathbf{B}-\mathbf{C} \geq 0$ and $\mathbf{E_{HH}} - \mathbf{A}^2 \geq 0$, where both will be strict as long as own and friends' characteristics are not perfectly correlated. Therefore, we get the following two results:

[1] The inconsistency of $\hat{\alpha}_{RF,1}^{cens,k}$ is in the same direction as the correlation between own and links' characteristics. If there is homophily on the characteristic, then $\operatorname{plim} \hat{\alpha}_1$ leads to inflation of the estimated own effect ($\operatorname{plim} \hat{\alpha}_1$ is of the same sign but larger in magnitude than true α_1).

[2] The inconsistency of $\hat{\alpha}_{RF,2}^{cens,k}$ is in the opposite direction of β_3. That is, it leads to attenuation.

While the direction is similar, this result is quite different from classical measurement error (see, e.g. Bound, Brown and Mathiowetz, 2001). Unlike in the classical case, the measurement error here may be correlated with observed variables. For example, if x_i is a binary variable that is randomly assigned, then measurement error will necessarily be negatively correlated with the observed (mismeasured) $\bar{x}_i$. Further, the assumption of order irrelevance gives us information about the inconsistency due to censoring, which can be used for correction as discussed in Section 5.

4. EMPIRICAL DEMONSTRATION OF INCONSISTENCY

4.1. *Sub-Censoring*

Now, I turn to empirical results. To demonstrate the effect of censoring on peer effects estimates, I adopt an approach that I refer to as *sub-censoring*. That is, given censoring induced by the survey design that allows respondents to name up to five links, I further censor the data as if we had only observed 1, 2, 3, or 4 links. With data thus sub-censored, I then construct peer group means $\bar{x}_{is}^{cens,k}$ for $k = 1, ..., 5$ and in turn estimate $\hat{\alpha}_{RF}^{cens,k}$ for each value of k. I perform this analysis with two datasets: data from the randomized trial in Cai, de Janvry and Sadoulet (2015) and the AddHealth Wave 1 in-school data (Harris, 2009).[12]

4.2. *China Insurance Results*

In a study of social determinants of insurance take-up in rural China, Cai, de Janvry and Sadoulet (2015) implemented a randomized experiment. To investigate the role of censoring, I focus on the specifications in Table 2, Column (2) of their original study. These results relate takeup for those assigned to Second Round sessions as a function of the percent of their friends who were assigned to First Round Intensive sessions, where assignment was random. This specification is given in Equation (4), where y_{is} is an indicator for insurance take-up for individual i in village s.

$$y_{is} = \alpha_0 + \alpha_1 Treat_{is} + \alpha_2 \overline{Treat}_{is} + Z'_{is}\gamma + \epsilon_{is} \tag{4}$$

In this context, $Treat_{is}$ is an indicator for individual i in village s being invited to an intensive information session, $\overline{Treat}_{is}$ is the fraction of individual i's friends who were assigned to the relevant treatment, and z_{is} is additional controls including village fixed effects. Performing the sub-censoring exercise requires constructing $\overline{Treat}_{is}$ when links are sub-censored at $k = 1$, 2, 3, and 4 links. That is, I calculate these means while dropping all but the first k links.[13], then estimate the parameters

[12]Sojourner (2013) performs a similar exercise on simulated missing data, the results of which are given in that paper's Figure 1.

[13]In this exercise, I employ the ordering of links provided in the publicly-available data, but I note that this link data was elicited in an unordered manner. Thanks to the authors for pointing this out. As confirmed in correspondence with the authors, the friendship data used in Cai, de Janvry and Sadoulet (2015) is essentially unordered in the sense that agents do not name their closest friend first, next closest second, etc. As an additional check on the importance of ordering, I perform the sub-censoring analysis dropping random links in Appendix C, with results analogous to those in Table 1 presented in Appendix Table A.2.

by regression with the censored means.

I perform the sub-censoring exercise using both "OUT" and "OR" definitions of network links, where the former is the definition that is used by the original authors.[14] The main results are in Table 1, where Panel A corresponds to the specification in Column (2) of the original authors' Table 2, which employs an "OUT" network definition. The estimates in the last column, in which up to five links are included in calculating $\overline{Treat}_i$, are the exact same estimates presented by Cai, de Janvry and Sadoulet (2015).

For both specifications, these results show the attenuation pattern under order irrelevance. Crucially, $\hat{\alpha}_2$, the coefficient on $\overline{Treat}_i$, the mismeasured peer mean variable, trends away from zero as we observe more and more links, suggesting that censoring here implies an underestimate of the true effect of average peers' treatment status.

TABLE 1

China Insurance Regression Results (Sub-Censored)

Max Number of Links	1	2	3	4	5
Panel A: "OUT" Network					
$Treat_{is}$ ($\hat{\alpha}_1$)	0.028	0.032	0.030	0.029	0.030
	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)
$\overline{Treat}_{is}$ ($\hat{\alpha}_2$)	0.115***	0.167***	0.182***	0.199***	0.291***
	(0.041)	(0.053)	(0.068)	(0.072)	(0.082)
R-squared	0.114	0.114	0.113	0.114	0.119
Panel B: "OR" Network					
$Treat_{is}$ ($\hat{\alpha}_1$)	0.028	0.028	0.028	0.029	0.031
	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)
$\overline{Treat}_{is}$ ($\hat{\alpha}_2$)	0.048	0.141*	0.251**	0.274***	0.311**
	(0.054)	(0.077)	(0.097)	(0.103)	(0.123)
R-squared	0.109	0.111	0.115	0.113	0.113

Notes: N = 1,255 in all specifications. All estimates correspond to specifications including village fixed effects and other controls as in Column (2) of Table 2 of Cai, de Janvry and Sadoulet (2015). Standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1.

Further, as predicted by Corollary 1, $\hat{\alpha}_1$, the estimated coefficient on $Treat_{is}$ (one's own treatment status), is unaffected by censoring. This conforms with the discussion in Section 3. That is, since treatment is randomly assigned at the individual level, the mean of any agent's peer group's treatment status should be independent of own treatment status.

[14] Under the "OUT" definition, i is linked to j if i identifies j as a link. Under the "OR" definition, i is linked to j if *either* i lists j or j lists i. For sub-censored networks, for a given k, an "OUT" link exits if i lists j among their first k links, while an "OR" link exists if either i lists j among their first k links or j lists i among their first k links.

To provide more evidence of the importance of censoring, I perform tests of equality of coefficients across different censoring rules k. The ideal analysis would test $\alpha^{cens,k} = \alpha$, where the latter would be estimated from uncensored data. However, we do not observe uncensored data. Accordingly, as a second-best alternative, for $k < 5$, I test whether the censored estimate is different than the (less) censored estimate generated with the full data (corresponding to $k = 5$). Thus, I test $\alpha_1^{cens,k} = \alpha_1^{cens,5}$ and $\alpha_2^{cens,k} = \alpha_2^{cens,5}$ for each $k < 5$. These results are given in Table A.4, where the final column gives results for a test of equality for all $k = 1, ..., 5$.

Under both network definitions, we fail to reject the null that $\alpha_1^{k,cens} = \alpha_1^{5,cens}$ for every $k < 5$, consistent with the result that $\text{plim}\, \hat{\alpha}^{cens,k} = \alpha_1$ regardless of censoring due to randomization. In contrast, we strongly reject the null that $\alpha_2^{cens,k} = \alpha_2^{cens,5}$ for $k < 3$ in most specifications, although the results are stronger for the "OUT" network in Panel A. This provides strong evidence that different censoring rules may lead to significantly different estimates of the peer effect parameter α_2.

4.3. *AddHealth*

The AddHealth data has been employed to study the associations between peers and a wide variety of academic and behavioral outcomes. For the purpose of this empirical exercise, I look at three academic and three behavioral outcomes from the Wave 1 in-school survey. These are: GPA in All Subjects, English, and Math; and whether the individual has Smoked, Drunk Alcohol, or Got Drunk in the past year. These variables are summarized in Appendix Table A.5, Panel A, while Panel BB summarizes the ten right-hand side variables that I use in the analysis, corresponding to x_{is}.

To demonstrate the impact of censoring, in the AddHealth data, I estimate specifications of the form

$$y_{is} = \alpha_0 + x'_{is}\alpha_1 + \overline{x}'_{is}\alpha_2 + \gamma_s + \epsilon_{is} \tag{5}$$

where y_{is} is one of the academic or non-academic outcomes, and x_{is} is a 10×1 vector of individual characteristics. For this sub-censoring exercise, I separately estimate the parameters after constructing sub-censored $\overline{x}_{is}$ for each $k = 1, ..., 5$. In the main text, I construct means using a

(symmetric) "OR" definition of friends, whereby two individuals are considered friends if either names the other among their first k nominations. In all specifications, school fixed effects (γ_s) are included to control for correlated effects.

TABLE 2
AddHealth Regression Results (Sub-Censored)

Censoring Rule (k)	1	2	3	4	5
Panel A: "OR" Network					
Age_{is}	-0.156***	-0.149***	-0.145***	-0.140***	-0.137***
	(0.012)	(0.011)	(0.011)	(0.011)	(0.011)
$Grade_{is}$	0.169***	0.158***	0.149***	0.148***	0.143***
	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)
$\overline{Age}_{is}$	-0.091***	-0.147***	-0.191***	-0.219***	-0.236***
	(0.011)	(0.017)	(0.020)	(0.024)	(0.025)
$\overline{Grade}_{is}$	0.079***	0.138***	0.186***	0.207***	0.227***
	(0.012)	(0.017)	(0.020)	(0.024)	(0.026)
R-squared	0.948	0.948	0.949	0.949	0.949
Panel B: "OUT" Network					
Age_{is}	-0.164***	-0.156***	-0.152***	-0.149***	-0.147***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
$Grade_{is}$	0.175***	0.171***	0.166***	0.166***	0.164***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
$\overline{Age}_{is}$	-0.083***	-0.137***	-0.177***	-0.195***	-0.209***
	(0.010)	(0.015)	(0.018)	(0.020)	(0.021)
$\overline{Grade}_{is}$	0.071***	0.117***	0.156***	0.170***	0.183***
	(0.012)	(0.018)	(0.019)	(0.021)	(0.021)
R-squared	0.949	0.950	0.950	0.950	0.950

Dependent Variable: GPA in All Subjects. N = 32,156 in all specifications in Panel A; N = 26,465 in all specifications in Panel B. Sample restricted to observations with non-missing data for all k. Standard errors in parentheses, clustered by school. Coefficients for other covariates in Table A.5, Panel B not shown. School fixed effects included in all specifications. *** p<0.01, ** p<0.05, * p<0.1.

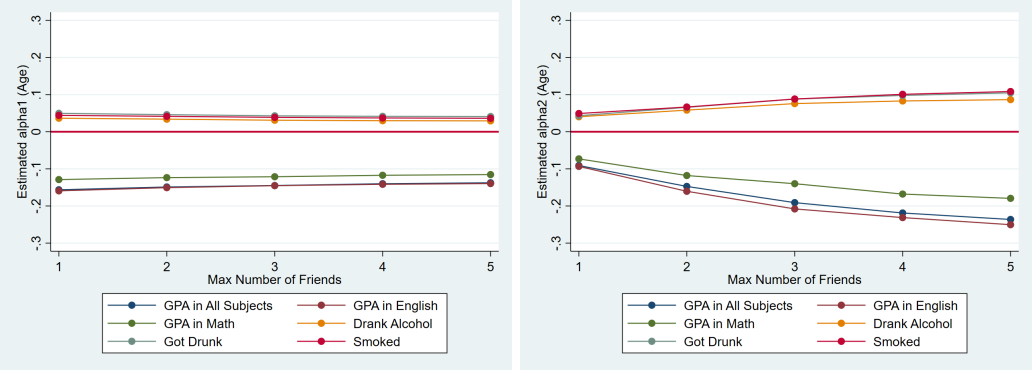
Table 2 gives the main results for outcome GPA in All Subjects. To focus the analysis, I present only estimates of the coefficients for two variables and means: Age and Grade. These results show clear patterns that are consistent with the analytic results under order irrelevance. Estimates for the effect of pees means ($\overline{Age}_{is}$, $\overline{Grade}_{is}$) are more attenuated (smaller in magnitude) with fewer observed friends. Consistent with homophily, estimates of the effect of own characteristics (Age_{is}, $Grade_{is}$) shift in the opposite direction from the effects on peer means for those same covariates. These results hold for both symmetric "OR" networks (Panel A) and asymmetric "OUT" networks (Panel B).[15]

[15]Similar to the results for the China Insurance application, I also perform this analysis after reversing the order of nominated friends, as an informal check on whether the patterns in Table 2 are due to listing friends in a particular

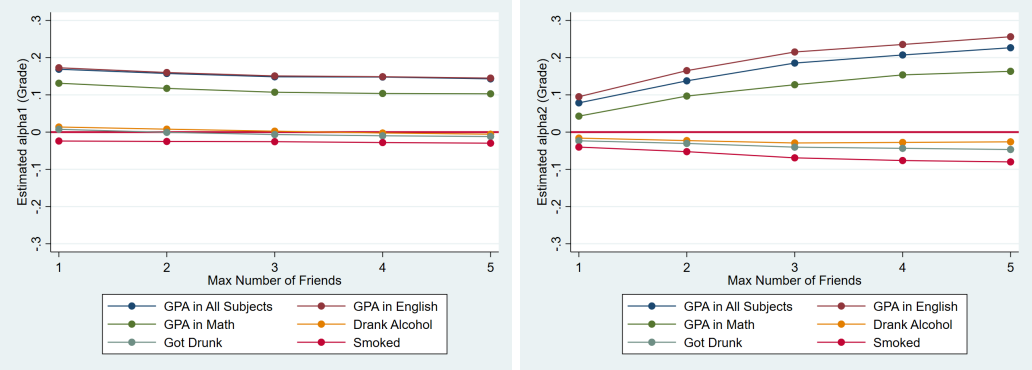
Figure 1 presents estimated coefficients on these same covariates for all six outcomes as k changes. the patterns here are clear: estimates of the effect of peer means trend away from zero as more friends are observed, while estimates of the effect of one's own characteristics (Female, Age) trend in the opposite direction. Appendix Figure A.1 shows that these results hold for asymmetric network definitions as well.

Figure 1: AddHealth Estimated Coefficients (Sub-Censored, "OR" Network)

(a) Coefficients for Age_{is} and $\overline{Age}_{is}$



(b) Coefficients for $Grade_{is}$ and $\overline{Grade}_{is}$



As with the China Insurance data, I also test for differences in estimates across different levels of censoring. Since I do not observe uncensored data and thus cannot directly estimate α, I test differences between the censored estimates and those that are *least* censored, corresponding to $k = 5$. These tests, therefore, calculate whether there are significant differences between estimates censored at $k < 5$ and those censored at $k = 5$. These results are given in Appendix Tables A.6 and A.7. Panel A gives results of joint tests of equality of own characteristics; Panel B gives results

order. These results are presented in Appendix Table A.3 and discussed in Appendix C.

of joint tests of equality of peer means; and Panel C simultaneously tests equality of own and peer characteristics.[16] From these results, we see that tests consistently reject equality, even when testing $k = 4$ vs. $k = 5$. Further, as shown by the results in the final column, we strongly reject equality across all values of k for all sets of regressors across all six outcomes. In sum, these results strongly support the claim that censoring of network data substantially affects estimates of parameters of linear-in-means models.

5. BIAS CORRECTION

Under some conditions, we can use the expression derived in Proposition 3 to derive a consistent estimate of α. Imporantly, this "cure" rests on the assumptions of order irrelevance (Assumption 3) as well as $\beta_1 = 0$. Additionaly, if there is reason to assume that $\mathbf{A} = \mathbf{C} = 0$, such as when the covariate is randomly assigned, then a simpler strategy proceeds from Corollary 3.

Here I provide a description of the correction method as well as demonstrating its performance in both datasets. Since we do not observe "true" degree in either dataset, the corrections begin with the sub-censored estimates in the previous section and proceed *as if* the full data—where $k = 5$—were indeed the true network. I also note that, when this is unknown, a bounding exercise may be used to provide bounds on the true parameter values.

5.1. *Method Description*

To derive the estimator, first I rewrite the result in Proposition 3 as $\text{plim}\, \hat{\alpha}^{cens,k} = \mathbf{B_k}\alpha$. In this expression,

$$\mathbf{B_k} = \begin{bmatrix} \mathbf{I} & z\mathbf{E_{xx}}^{-1}\mathbf{A}\mathbf{D}^{-1}(\mathbf{B}-\mathbf{C}) \\ \mathbf{0} & \mathbf{I} - z\mathbf{D}^{-1}(\mathbf{B}-\mathbf{C}) \end{bmatrix}, \text{where } \mathbf{D} = \mathbf{C} + \mathbb{E}[\frac{1}{d^{cens,k}}](\mathbf{B}-\mathbf{C}) - \mathbf{A'}\mathbf{E_{xx}}^{-1}\mathbf{A} \tag{6}$$

$$z = \mathbb{E}[\frac{1}{d^{cens,k}}] - \mathbb{E}[\frac{1}{d^{true}}]$$

[16] In performing these tests, it is crucial to account for covariance in the estimators across specifications. To see this, recall that $\hat{\alpha}_{RF}^{cens,k}$ is simply a regression of the outcome y_{is} on a constant, x_i, and a censored version of mean $\bar{x}_{is}^{cens,k}$. The dependent variable and the first two regressors are constant across different censoring rules, and $\bar{x}_{is}^{cens,k}$—the mismeasured peer mean—is highly correlated across k. Accordingly, $\hat{\alpha}_{RF}^{cens,k}$ are likely to be highly correlated across k, and simply comparing differences in point estimates to standard errors in Table 2 may be highly misleading. Therefore, the test results presented in Tables A.6, which are derived from the same point estimates as in Table 2 while accounting for covariance across specifications, are not inconsistent with the regression results presented in the latter table. Mechanically, these tests are calculated as post-estimation tests from multi-equation GMM, where each of five equations in the GMM system corresponds to estimates censored at $k = 1, ..., 5$, and clustering by school allows residuals to be arbitrarily correlated across equations within schools.

Consistent estimation of α requires $\hat{\mathbf{B}}_\mathbf{k}$, a consistent estimate of $\mathbf{B}_\mathbf{k}$, which allows for construction of a consistent estimate of α as $\hat{\mathbf{B}}_\mathbf{k}^{-1}\hat{\alpha}^{cens,k}$.

Under order irrelevance, we can construct consistent estimates of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{E_{xx}}$, and $\mathbb{E}[\frac{1}{d^{cens,k}}]$ directly from the censored data. Some particular estimators, corresponding to empirical analogues of the target estimands, are provided in Appendix B.[17] That leaves only unobserved $\mathbb{E}[\frac{1}{d^{true}}]$. Consistent estimation of this parameter, and thus of α, requires supplemental data on the degree distribution. Appendix B briefly discusses how an estimate may be derived from either Aggregate Relational Data or two-step edgewise sampling, but in principle other methods may be used.

Next, a simpler form of bias correction is possible when the covariate $\mathbf{x}$ has been randomly assigned, such as is common in many randomized trials. In this case, the probability limit of $\hat{\alpha}^{cens,k}$ is given by Corollary 3, and

$$\mathbf{B}_\mathbf{k} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \frac{\mathbb{E}[\frac{1}{d^{true}}]}{\mathbb{E}[\frac{1}{d^{cens,k}}]}\mathbf{I} \end{bmatrix} \tag{7}$$

Therefore, when $\mathbf{A} = \mathbf{C} = 0$, bias correction only depends on consistent estimates of $\mathbb{E}[\frac{1}{d^{cens,k}}]$ (from the censored data) and $\mathbb{E}[\frac{1}{d^{true}}]$ (from supplemental data on the degree distribution). In results below, this simpler strategy is referred to as the "Restricted Method," in contrast to the "Full Method."

Finally, in cases where supplemental data is unavailable, bounding is possible. To see this, let $p^{cens,k} \in [0,1]$ be the probability that degree is censored for a given k, and consider the following decomposition:

$$\mathbb{E}[\frac{1}{d^{true}}] = p^{cens,k}\mathbb{E}[\frac{1}{d^{true}}|\text{censored}] + (1-p^{cens,k})\mathbb{E}[\frac{1}{d^{true}}|\text{not censored}] \tag{8}$$

Since $\frac{1}{d^{true}} > 0$ and censored degree coincides with true degree for uncensored agents, this becomes

$$\mathbb{E}[\frac{1}{d^{true}}] > \underbrace{(1-p^{cens,k})}_{\substack{\text{fraction} \\ \text{not censored}}} \underbrace{\mathbb{E}[\frac{1}{d^{cens,k}}|\text{not censored}]}_{\substack{\text{mean among} \\ \text{uncensored}}} \tag{9}$$

[17] Note, however, that we need to observe at least two links for some agents in order to consistently estimate $\mathbf{C}$.

Therefore, the "tightness" of this bound depends both on the fraction censored and on the average inverse degree among the uncensored. Importantly, this bound can be estimated from the censored degree distribution.

5.2. *Bias Correction in China Insurance Data*

Here, I present evidence of the performance of bias correction in the China Insurance data. Note that the assumption of order irrelevance is particularly plausible in this setting: since treatment is assigned randomly, the treatment assignment of all links follows the same distribution, regardless of the order they are listed in, and regardless of whether or not they are censored.

These results are given in Table 3. Panel A gives results using the Full Method, which estimates sample analogues of $\mathbf{B_k}$ as given in Equation (6). From this, we see that estimates of the effect of own treatment $Treat_{is}$ are again inesensitive to censoring. In contrast to the uncorrected results in Table 1, however, estimates of the effects of peers' average treatment $\overline{Treat}_{is}$ are also relatively insensitive to censoring. Results of the Restricted Method are given in Panel B, and show similar results: estimates do not show the attenuation patterns found in the uncorrected results. Results for "OR" network are qualitatively similar, and are presented in Appendix Table A.8. In sum, under both network definitions, the corrected coefficient estimates are insensitive to censoring, in sharp contrast to the uncorrected estimates as given in Table 1.

TABLE 3
China Insurance Corrected Estimates ("OUT" Network)

Max Number of Links	2	3	4	5
Panel A: Full Method				
$Treat_{is}$ ($\hat{\alpha}_1$)	0.034	0.033	0.030	0.030
	(0.033)	(0.033)	(0.033)	(0.033)
$\overline{Treat}_{is}$ ($\hat{\alpha}_2$)	0.339**	0.301***	0.281***	0.291***
	(0.152)	(0.098)	(0.088)	(0.083)
Panel B: Restricted Method				
$Treat_{is}$ ($\hat{\alpha}_1$)	0.031	0.032	0.030	0.030
	(0.033)	(0.033)	(0.033)	(0.033)
$\overline{Treat}_{is}$ ($\hat{\alpha}_2$)	0.318**	0.327***	0.293***	0.291***
	(0.141)	(0.104)	(0.091)	(0.083)

Notes: N = 1,255 in all specifications. All estimates correspond to specifications including village fixed effects and other controls as in Column (2) of Table 2 of Cai, de Janvry and Sadoulet (2015). "OUT" network definition as used by original authors. Bootstrap standard errors in parentheses calculated from 1000 repetitions, sampled by village. *** p<0.01, ** p<0.05, * p<0.1.

5.3. *Bias Correction in AddHealth*

I perform a similar exercise to correct the AddHealth estimates. Since the data clearly shows that networks are not independent of covariates $\mathbf{x}$, the Restricted Method is inappropriate in this setting, so I only present results for the Full Method in AddHealth. In this data, an additional complication arises, however, due to the fact that there is a substantial quantity of missing peer data. Consider, for example, the sequences of friendship nominations for Students 1 and 2 as detailed in Table 4. Some friendship nominations are matched ("M" in the table) and others are not matched ("NM" in the table), where the latter category include nominations of students in other schools, outside the school, or missing for some other reason. Missing values in the table, denoted ".", indicate that a student did not nominate that many friends.

TABLE 4
AddHealth Missingness Example

	Male Friends					Female Friends				
Student	1	2	3	4	5	1	2	3	4	5
1	M	NM	NM	M	.	M	NM	.	.	.
2	NM	NM	M	.	.	M	NM	M	M	NM

If we were to naively construct a friendship mean for student 1, the most natural procedure is to drop the missing nominations and construct a mean from the remaining ones. In truth, Student 1 has nominated 6 friends (4 male, 2 female), but this would create a mean of only the three who are not missing in the data.

Order irrelevance, however, implies a "fix" for this problem in addition to censoring. Intuitively, the problem of missing friends' covariates is similar to missing friend data due to censoring. Both issues imply that we only observe a subset of each individual's peers' characteristics. Under the assumption of order irrelevance, the same method can be employed to correct estimates due to both issues. Importantly, the data gives us the number of friends with missing data, which is analogous to knowing the degree distribution with censored data.

Table 5 gives estimates for the outcome GPA in All Subjects, the same outcome investigated in Table 2. In contrast to the estimates presented there, however, here we do not see a clear pattern as we observe more links (as k goes from 2 to 5). Rather, it appears that estimates are relatively insensitive to k.

TABLE 5

AddHealth Corrected Results

Censoring Rule (k)	2	3	4	5
Panel A: "OR" Network				
Age_{is}	-0.122***	-0.126***	-0.125***	-0.126***
	(0.011)	(0.010)	(0.010)	(0.010)
$Grade_{is}$	0.129***	0.126***	0.134***	0.131***
	(0.020)	(0.019)	(0.018)	(0.017)
$\overline{Age}_{is}$	-0.276***	-0.293***	-0.297***	-0.297***
	(0.040)	(0.036)	(0.036)	(0.035)
$\overline{Grade}_{is}$	0.262***	0.286***	0.280***	0.284***
	(0.039)	(0.035)	(0.035)	(0.035)
Panel B: "OUT" Network				
Age_{is}	-0.129***	-0.132***	-0.132***	-0.133***
	(0.011)	(0.011)	(0.011)	(0.011)
$Grade_{is}$	0.160***	0.148***	0.156***	0.155***
	(0.019)	(0.017)	(0.016)	(0.015)
$\overline{Age}_{is}$	-0.270***	-0.287***	-0.277***	-0.277***
	(0.038)	(0.034)	(0.032)	(0.031)
$\overline{Grade}_{is}$	0.225***	0.258***	0.240***	0.242***
	(0.044)	(0.035)	(0.034)	(0.031)

Dependent Variable: GPA in All Subjects. N = 32,156 in all specifications in Panel A; N = 26,465 in all specifications in Panel B. Sample restricted to observations with non-missing data for all k. Bootstrap standard errors in parentheses, sampled by school. School fixed effects included in all specifications.Coefficients for other covariates in Table A.5, Panel B not shown. *** p<0.01, ** p<0.05, * p<0.1.

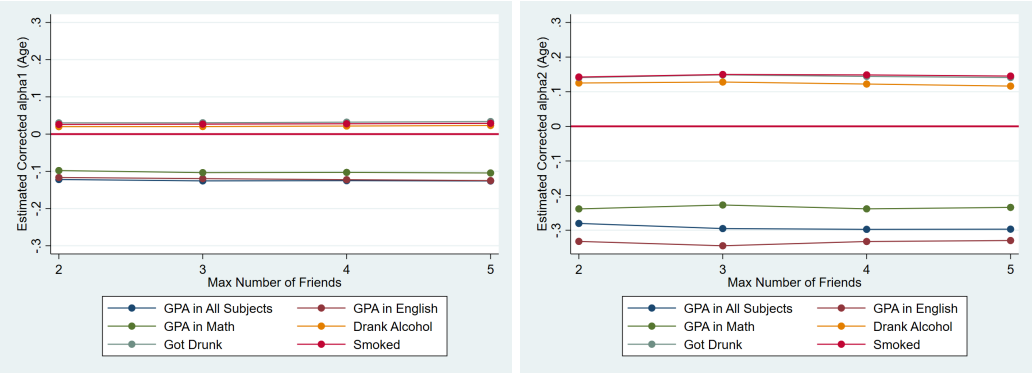
Corrected coefficient estimates, for the "OR" network definition, for these same covariates with all six outcomes are presented in Figure 2. In sharp contrast to the uncorrected results presented in Figure 1, estimated parameters do not exhibit clear trends as k increases. These same patterns are seen using asymmetric "OUT" networks, as shown in Appendix Figure A.2. In sum, in contrast to the uncorrected estimates, the corrected estimates are relatively insensitive to the censoring rule across a large number of outcomes, covariates, and network definitions.

6. DISCUSSION

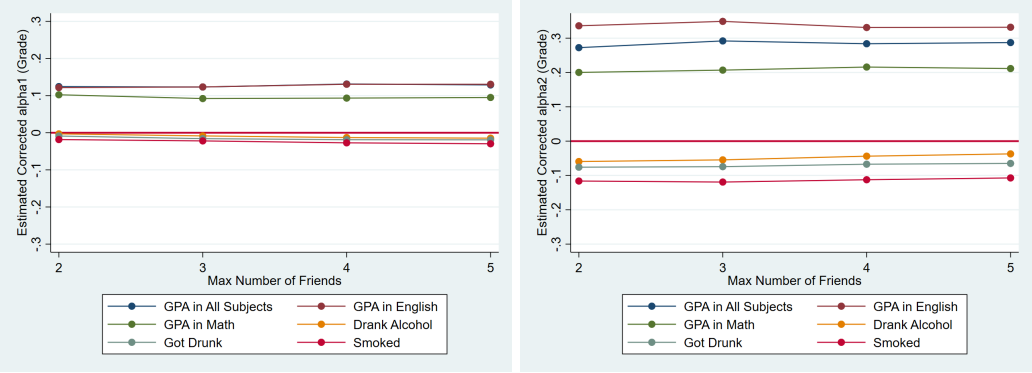
Censoring is a common feature of network datasets. While censoring has been noted as a potential issue by a number of authors, the relationship between the widespread data-collection technique that generates censoring and estimates of peer effects parameters has received little attention in the literature. I fill this gap by providing the first systematic treatment of the importance of censoring in network-based peer effects estimates. I show analytically that censoring implies inconsistent estimates; under order irrelevance, it leads to attenuation in peer effects estimates. Sub-censoring

Figure 2: AddHealth Corrected Coefficients ("OR" Networks)

(a) Coefficients for Age_{is} and $\overline{Age}_{is}$



(b) Coefficients for $Grade_{is}$ and $\overline{Grade}_{is}$



in the two empirical exercises here suggest that this attenuation can be quantitatively meaningful.

The analysis here provides two main contributions. First, for researchers dealing with censored data, the analytic and sub-censored results suggest that estimates of peer influence that employ censored network data systematically *under-estimate* its magnitude. In turn, this suggests that estimates in the literature may be too small, and that estimates derived from censored network data are best interpreted as lower bounds (in magnitude). Further, persistent attenuation implies that statistical tests for the existence of peer effects may fail to find peer effects when in fact such effects exist.

Second, I provide a correction method and conditions under which it will lead to consistent parameter estimates. The key assumption on the censored data is order irrelevance, which ensures that the distribution of peer mean variables is constant across orderings. The method also requires supplemental data from which to calculate $\mathbb{E}[\frac{1}{d^{true}}]$. An example of a network survey method that

ensures order irrelevance and allows for consistent estimation of this parameter is *two-step edge-wise sampling*, as was performed in Conley and Udry (2012).[18]

I conclude by noting that the results in this paper rely upon the strong assumption of network exogeneity, as detailed in Assumption 1. The fact that the empirical, sub-censored results in AddHealth show patterns in accord with the analytic results bolsters the credibility of those results, but the assumption of network exogeneity may still be violated in many applications. A growing series of analyses investigates methods to deal with network endogeneity in estimating parameters of linear-in-means models (Auerbach, 2019; Badev, 2017; Griffith, 2019; Johnsson and Moon, 2019). Investigation of the intersection between network censoring and network endogeneity is left for future research.

[18]It bears noting that this method also conforms to the requirements of Boucher and Houndetoungan (2019).

REFERENCES

Ammermueller, Andreas, and Jörn-Steffen Pischke. 2009. "Peer effects in European primary schools: Evidence from the progress in international reading literacy study." *Journal of Labor Economics*, 27(3): 315–348.

Auerbach, Eric. 2019. "Identification and Estimation of a Partially Linear Regression Model using Network Data." Unpublished Working Paper.

Badev, Anton. 2017. "Discrete Games in Endogenous Networks: Equilibria and Policy." Unpublished Working Paper.

Banerjee, Abhijit, Arun G. Chandrasekhar, Ester Duflo, and Matthew O. Jackson. 2012. "The Diffusion of Microfinance." NBER Working Paper No. 17743.

Baumann, Leonie. 2017. "A Model of Weighted Network Formation." Unpublished Working Paper.

Bloch, Francis, and Bhaskar Dutta. 2009. "Communication Networks with Endogenous Link Strength." *Games and Economic Behavior*, 66(1): 39–56.

Boucher, Vincent, and Aristide Houndetoungan. 2019. "Eestimating Peer Effects Using Partial Network Data." Unpublished Working Paper.

Boucher, Vincent, Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. 2014. "Do Peers Affect Student Achievement? Evidence from Canada Using Group Size Variation." *Journal of Applied Econometrics*, 29(1): 91–109.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement Error in Survey Data." In *Handbook of Econometrics*. Vol. 5, 3805–3843.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. 2009. "Identification of Peer Effects through Social Networks." *Journal of Econometrics*, 150(1): 41–55.

Breza, Emily, Arun G. Chandrasekhar, Tyler H. McCormick, and Mengjie Pan. 2020. "Using aggregated relational data to feasibly identify network structure without network data." *American Economic Review*, 110(8): 2454–84.

Caeyers, Bet, and Marcel Fafchamps. 2019. "Exclusion Bias in the Estimation of Peer Effects." Unpublished Working Paper.

Cai, Jing, Alain de Janvry, and Elisabeth Sadoulet. 2015. "Social Networks and the Decision

to Insure." *American Economic Journal: Applied Economics*, 7(2): 81–108.

Carrell, Scott, Bruce Sacerdote, and James West. 2013. "From Random Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation." *Econometrica*, 81(3): 855–882.

Chandrasekhar, Arun, and Randall Lewis. 2011. "Econometrics of Sampled Networks." Unpublished Working Paper.

Comola, Margherita, and Silvia Prina. 2018. "Treatment Effect Accounting for Network Changes: Evidence from a Randomized Intervention." Unpublished Working Paper.

Conley, Timothy G., and Christopher R. Udry. 2012. "Learning about a New Technology: Pineapple in Ghana." *American Economic Review*, 35–69.

Delavallade, Clara, Alan Griffith, and Rebecca Thornton. 2016. "Unintended Spillovers of a Girls' Empowerment Program in Rural Rajasthan: Targeting Matters." Unpublished Working Paper.

de Paula, Áureo, Imran Rasul, and Pedro Souza. 2018. "Recovering Social Networks from Panel Data: Identification, Simulations and an Application." Unpublished Working Paper.

Fosdick, Bailey K., and Peter D. Hoff. 2015. "Testing and Modeling Dependencies between a Network and Nodal Attributes." *Journal of the American Statistical Association*, 110(511): 1047–1056.

Graham, Bryan S. 2015. "Methods of Identification in Social Networks." *Annual Review of Economics*, 7(1): 465–485.

Griffith, Alan. 2019. "Random Assignment with Non-Random Peers: A Structural Approach to Counterfactual Treatment Assessment." Unpublished Working Paper.

Hardy, Morgan, Rachel Heath, Wesley Lee, and Tyler McCormick. 2019. ""Estimating spillovers using imprecisely measured networks." Unpublished Working Paper.

Harris, Kathleen Mullan. 2009. *The National Longitudinal Study of Adolescent to Adult Health (Add Health), Waves I and II, 1994-1996.* Carolina Population Center, University of North Carolina at Chapel Hill.

Hoff, Peter, Bailey Fosdick, Alex Volfovsky, and Katherine Stovel. 2013. "LIkelihoods for Fixed Rank Nomination Networks." *Network Science*, 1(3): 253–277.

Johnsson, Ida, and Hyungsik Roger Moon. 2019. "Estimation of peer effects in endogenous social networks: Control function approach." *Review of Economics and Statistics*, 1–51.

Kandpal, Eeshani, and Kathy Baylis. 2013. "Expanding Horizons: Can Women's Support Groups Diversify Peer Networks in Rural India?" *American Journal of Agricultural Economics*, 95(2): 360–367.

Leung, Michael. 2019. "A Weak Law of Moments for Pairwise-Stable Networks." *Journl of Econometrics*. Forthcoming.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies*, 60(3): 531–542.

McCormick, Tyler H, and Tian Zheng. 2015. "Latent surface models for networks using Aggregated Relational Data." *Journal of the American Statistical Association*, 110(512): 1684–1695.

Mele, Angelo. 2017. "A Structural Model of Dense Network Formation." *Econometrica*, 85(3): 825–850.

Ngatia, Muthoni. 2015. "Social Interactions and Individual Reproductive Decisions." Unpublished Working Paper.

Oster, Emily, and Rebecca Thornton. 2012. "Determinants of Technology Adoption: Peer Effects in Menstrual Cup Take-Up." *Journal of the European Economic Association*, 10(6): 1263–1293.

Sojourner, Aaron. 2013. "Identification of Peer Effects with Missing Peer Data: Evidence from Project STAR." *Economic Journal*, 123(569): 574–605.

Thirkettle, Matthew. 2019. "Identification and Estimation of Network Statistics with Missing Link Data." Unpublished Working Paper.

Tjernström, Emilia. 2017. "Learning from Others in Heterogeneous Environments." Unpublished Working Paper.

APPENDIX A: PROOFS OF PROPOSITIONS

Proposition 1

Begin with the definition $\hat{\alpha}_{RF} = ([\mathbf{x}, \mathbf{Gx}]'[\mathbf{x}, \mathbf{Gx}]')^{-1}[\mathbf{x}, \mathbf{Gx}]'\mathbf{y}$. The Slutzky Theorem provides that

$$\text{plim}\, \hat{\alpha}_{RF} = \text{plim}(\frac{1}{N}[\mathbf{x}, \mathbf{Gx}]'[\mathbf{x}, \mathbf{Gx}])^{-1}\, \text{plim}\left(\frac{1}{N}[\mathbf{x}, \mathbf{Gx}]'\mathbf{y}\right) \tag{A.1}$$

Independence across groups s and $s \to \infty$ implies

$$\text{plim}(\frac{1}{N}[\mathbf{x}, \mathbf{Gx}]'[\mathbf{x}, \mathbf{Gx}])^{-1} = (\mathbb{E}[\mathbf{x}, \mathbf{Gx}]'[\mathbf{x}, \mathbf{Gx}])^{-1} \tag{A.2}$$

Next, $|\beta_1| < 1$ implies that $(I - \beta_1\mathbf{G})$ is invertible. Therefore, Equation (2) can be rearranged as

$$\mathbf{y} = (I - \beta_1\mathbf{G})^{-1}(\mathbf{x}\beta_2 + \mathbf{Gx}\beta_3 + \epsilon) \tag{A.3}$$

Further, $|\beta_1| < 1$ allows for the expansion $(I - \beta_1\mathbf{G})^{-1} = \sum_{r=0}^{\infty} \beta_1^r\mathbf{G}^r$ (see, e.g. Bramoullé, Djebbari and Fortin, 2009; de Paula, Rasul and Souza, 2018). Therefore,

$$\mathbf{y} = (I + \beta_1\mathbf{G} + \beta_1^2\mathbf{G}^2 + ...)(\mathbf{x}\beta_2 + \mathbf{Gx}\beta_3 + \epsilon) \tag{A.4}$$

Rewritten in more convenient matrix form, this becomes

$$\mathbf{y} = \mathbf{x}\beta_2 + \mathbf{Gx}(\beta_3 + \beta_1\beta_2) + \sum_{r=1}^{\infty} \mathbf{G^{r+1}x}\beta_1^r(\beta_3 + \beta_1\beta_2) + \sum_{r=0}^{\infty} \beta_1^r\mathbf{G}^r\epsilon \tag{A.5}$$

$$= \left[\mathbf{x}, \mathbf{Gx}\right]\begin{bmatrix}\beta_2 \\ (\beta_3 + \beta_1\beta_2)\end{bmatrix} + \sum_{r=1}^{\infty} \mathbf{G^{r+1}x}\beta_1^k(\beta_3 + \beta_1\beta_2) + \sum_{r=0}^{\infty} \beta_1^k\mathbf{G}^k\epsilon \tag{A.6}$$

and thus

$$\left[\mathbf{x}, \mathbf{Gx}\right]'\mathbf{y} = \left[\mathbf{x}, \mathbf{Gx}\right]'\left[\mathbf{x}, \mathbf{Gx}\right]\begin{bmatrix}\beta_2 \\ (\beta_3 + \beta_1\beta_2)\end{bmatrix} + \left[\mathbf{x}, \mathbf{Gx}\right]'\sum_{r=1}^{\infty} \mathbf{G^{r+1}x}\beta_1^k(\beta_3 + \beta_1\beta_2)$$
$$+ \left[\mathbf{x}, \mathbf{Gx}\right]'\sum_{r=0}^{\infty} \beta_1^r\mathbf{G}^r\epsilon \tag{A.7}$$

Multiply by $\frac{1}{N}$ then take probability limits, yielding

$$\text{plim}\left(\frac{1}{N}[\mathbf{x}, \mathbf{Gx}]'\mathbf{y}\right) = \mathbb{E}[[\mathbf{x}, \mathbf{Gx}]'[\mathbf{x}, \mathbf{Gx}]]\begin{bmatrix}\beta_2 \\ (\beta_3 + \beta_1\beta_2)\end{bmatrix} + \sum_{r=1}^{\infty} \beta_1^k\begin{bmatrix}\mathbb{E}[\mathbf{x'G^{r+1}x}] \\ \mathbb{E}[\mathbf{x'G'G^{r+1}x}]\end{bmatrix}(\beta_3 + \beta_1\beta_2)$$
$$+ \mathbb{E}[[\mathbf{x}, \mathbf{Gx}]'\sum_{r=0}^{\infty} \beta_1^k\mathbf{G}^k\epsilon] \tag{A.8}$$

Assumption 1 implies that the final term is zero. Therefore,

$$\text{plim}\left(\frac{1}{N}[\mathbf{x}, \mathbf{Gx}]'\mathbf{y}\right) = \mathbb{E}[[\mathbf{x}, \mathbf{Gx}]'[\mathbf{x}, \mathbf{Gx}]]\begin{bmatrix}\beta_2 \\ (\beta_3 + \beta_1\beta_2)\end{bmatrix} + \sum_{r=1}^{\infty} \beta_1^k\begin{bmatrix}\mathbb{E}[\mathbf{x'G^{r+1}x}] \\ \mathbb{E}[\mathbf{x'G'G^{r+1}x}]\end{bmatrix}(\beta_3 + \beta_1\beta_2) \tag{A.9}$$

Substitute Equations (A.2) and (A.9) into Equation (A.1) to give

$$\text{plim}\, \hat{\alpha}_{RF} = \begin{bmatrix}\beta_2 \\ (\beta_3 + \beta_1\beta_2)\end{bmatrix} + (\mathbb{E}[\mathbf{x}, \mathbf{Gx}]'[\mathbf{x}, \mathbf{Gx}])^{-1}\sum_{r=1}^{\infty} \beta_1^k\begin{bmatrix}\mathbb{E}[\mathbf{x'G^{r+1}x}] \\ \mathbb{E}[\mathbf{x'G'G^{r+1}x}]\end{bmatrix}(\beta_3 + \beta_1\beta_2) \tag{A.10}$$

which can be rewritten as the result

$$\text{(A.11)}\quad \operatorname{plim}\hat{\alpha}_{RF} = \begin{bmatrix} \beta_2 \\ (\beta_3+\beta_1\beta_2) \end{bmatrix} + \begin{bmatrix} \mathbf{E_{xx}} & \mathbf{E_{xG}}' \\ \mathbf{E_{xG}} & \mathbf{E_{GG}} \end{bmatrix}^{-1} \sum_{r=1}^{\infty} \beta_1^k \begin{bmatrix} \mathbb{E}[\mathbf{x}'\mathbf{G}^{\mathbf{r+1}}\mathbf{x}] \\ \mathbb{E}[\mathbf{x}'\mathbf{G}'\mathbf{G}^{\mathbf{r+1}}\mathbf{x}] \end{bmatrix} (\beta_3+\beta_1\beta_2)$$

Proposition 2

First, $|\beta|<1$ allows for the expansion $(I-\beta_1\mathbf{G})^{-1}=\sum_{r=0}^{\infty}\beta_1^k\mathbf{G}^k$. Therefore,

$$\text{(A.12)}\quad \mathbf{y} = \sum_{r=0}^{\infty}\beta_1^k\mathbf{G}^k(\mathbf{x}\beta_2+\mathbf{G}\mathbf{x}\beta_3+\epsilon) = [\mathbf{x},\mathbf{G}\mathbf{x}]\begin{bmatrix}\beta_2\\ \beta_1\beta_2+\beta_3\end{bmatrix} + \sum_{r=1}^{\infty}\beta_1^k\mathbf{G}^{r+1}\mathbf{x}(\beta_1\beta_2+\beta_3) + \sum_{r=0}^{\infty}\beta_1^k\mathbf{G}^k\epsilon$$

So,

$$\text{(A.13)}\quad [\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{y} = [\mathbf{x},\mathbf{H_k}\mathbf{x}]'[\mathbf{x},\mathbf{G}\mathbf{x}]\begin{bmatrix}\beta_2\\ \beta_1\beta_2+\beta_3\end{bmatrix} + \sum_{r=1}^{\infty}\beta_1^k[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{G}^{r+1}\mathbf{x}(\beta_1\beta_2+\beta_3) + \sum_{r=0}^{\infty}\beta_1^k[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{G}^k\epsilon$$

Multiplying by $\frac{1}{N}$ and taking the probability limit as $N\to\infty$ gives

$$\begin{aligned}\mathbb{E}[[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{y}] &= \mathbb{E}[[\mathbf{x},\mathbf{H_k}\mathbf{x}]'[\mathbf{x},\mathbf{G}\mathbf{x}]]\begin{bmatrix}\beta_2\\ \beta_1\beta_2+\beta_3\end{bmatrix} + \sum_{r=1}^{\infty}\beta_1^k\mathbb{E}[[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{G}^{r+1}\mathbf{x}](\beta_1\beta_2+\beta_3)\\ &\quad + \sum_{r=0}^{\infty}\beta_1^k\mathbb{E}[[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{G}^k\epsilon] \qquad \text{(A.14)}\end{aligned}$$

The final term is equal to zero due to Assumption 1 and thus

$$\text{(A.15)}\quad \mathbb{E}[[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{y}] = \mathbb{E}[[\mathbf{x},\mathbf{H_k}\mathbf{x}]'[\mathbf{x},\mathbf{G}\mathbf{x}]]\begin{bmatrix}\beta_2\\ \beta_1\beta_2+\beta_3\end{bmatrix} + \sum_{r=1}^{\infty}\beta_1^r\mathbb{E}[[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{G}^{r+1}\mathbf{x}](\beta_1\beta_2+\beta_3)$$

Next, for each $r\geq 1$,

$$\text{(A.16)}\quad \mathbb{E}[[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{G}^{r+1}\mathbf{x}] = \mathbf{C_{HG}}\mathbf{C_{GG}}^{-1}\mathbb{E}[\mathbf{x},\mathbf{G}\mathbf{x}]'\mathbf{G}^{r+1}\mathbf{x}] + \mathbf{C_{HG}}\left(\mathbf{C_{HG}}^{-1}\mathbb{E}[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{G}^{r+1}\mathbf{x}] - \mathbf{C_{GG}}^{-1}\mathbb{E}[\mathbf{x},\mathbf{G}\mathbf{x}]'\mathbf{G}^{r+1}\mathbf{x}]\right)$$

where $\mathbf{C_{GG}}$, $\mathbf{C_{HG}}$, and $\mathbf{C_{HH}}$ are defined in the text. Substitute (A.16) into (A.15) and collect terms, which gives

$$\begin{aligned}\mathbb{E}[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{y}] &= \mathbf{C_{HG}}\left(\begin{bmatrix}\beta_2\\ \beta_1\beta_2+\beta_3\end{bmatrix} + \sum_{r=1}^{\infty}\beta_1^k\mathbf{C_{GG}}^{-1}\mathbb{E}[\mathbf{x},\mathbf{G}\mathbf{x}]'\mathbf{G}^{r+1}\mathbf{x}](\beta_1\beta_2+\beta_3)\right)\\ &\quad + \sum_{r=1}^{\infty}\beta_1^k\mathbf{C_{HG}}\left(\mathbf{C_{HG}}]^{-1}\mathbb{E}[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{G}^{r+1}\mathbf{x}] - \mathbf{C_{GG}}^{-1}\mathbb{E}[\mathbf{x},\mathbf{G}\mathbf{x}]'\mathbf{G}^{r+1}\mathbf{x}])(\beta_1\beta_2+\beta_3)\right) \qquad \text{(A.17)}\end{aligned}$$

Substitute α from Proposition 1,

$$\text{(A.18)}\quad \mathbb{E}[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{y}] = \mathbf{C_{HG}}\left(\alpha + \sum_{r=1}^{\infty}\beta_1^k(\mathbf{C_{HG}}^{-1}\mathbb{E}[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{G}^{r+1}\mathbf{x}] - \mathbf{C_{GG}}^{-1}\mathbb{E}[\mathbf{x},\mathbf{G}\mathbf{x}]'\mathbf{G}^{r+1}\mathbf{x})(\beta_1\beta_2+\beta_3)]\right)$$

Finally, by the Slutzky Theorem,

$$\text{(A.19)}\quad \operatorname{plim}\hat{\alpha}_{RF}^{cens,k} = (\operatorname{plim}[\mathbf{x},\mathbf{H_k}\mathbf{x}]'[\mathbf{x},\mathbf{H_k}\mathbf{x}])^{-1}(\operatorname{plim}[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{y}) = \mathbf{C_{HH}}^{-1}\mathbb{E}[[\mathbf{x},\mathbf{H_k}\mathbf{x}]'\mathbf{y}]$$

Therefore, substitute (A.18) into (A.19), yielding

$$\text{(A.20)} \quad \operatorname{plim} \hat{\alpha}_{RF}^{cens,k} = \mathbf{C_{HH}}^{-1}\mathbf{C_{HG}} \left(\alpha + \sum_{r=1}^{\infty} \beta_1^r (\mathbf{C_{HG}}^{-1}\mathbb{E}[\mathbf{x}, \mathbf{H_k x}]'\mathbf{G}^{r+1}\mathbf{x}] - \mathbf{C_{GG}}^{-1}\mathbb{E}[\mathbf{x}, \mathbf{G x}]'\mathbf{G}^{r+1}\mathbf{x}])(\beta_1\beta_2 + \beta_3) \right)$$

Proposition 3

I proceed in three steps. The first two derive formulas for $\mathbf{E_{xH}}$, $\mathbf{E_{xG}}$, $\mathbf{E_{HG}}$, and $\mathbf{E_{HH}}$. Step 3 assembles these pieces together to derive the final expression.

Step 1: $\mathbf{E_{xH}} = \mathbf{E_{xG}} = \mathbf{A}$

First, decompose the expectation by true degree $\overline{d}$ and and censored degree $d^{cens,k}$ as follows:

$$\text{(A.21)} \quad \mathbf{E_{xG}} = \sum_{d^{cens,k}} \sum_{\overline{d} \geq d^{cens,k}} p(d^{cens,k}, \overline{d}) \mathbb{E}[\mathbf{x}'\mathbf{G}\mathbf{x} | d^{cens,k}, \overline{d}]$$

$$\text{(A.22)} \quad = \sum_{d^{cens,k}} \sum_{\overline{d} \geq d^{cens,k}} p(d^{cens,k}, \overline{d}) \mathbb{E}[x_i (\frac{1}{\overline{d}} \sum_{v=1}^{\overline{d}} \bar{x}'_{i(v)}) | d^{cens,k}, \overline{d}]$$

Next, rearrange and apply order irrelevance as follows:

$$\text{(A.23)} \quad \mathbb{E}[x_i (\frac{1}{\overline{d}} \sum_{v=1}^{\overline{d}} \bar{x}'_{i(v)}) | d^{cens,k}, \overline{d}] = \frac{1}{\overline{d}} \overline{d} \mathbb{E}[x_i \bar{x}'_{i(v)} | d^{cens,k}, \overline{d}] = \mathbf{A}$$

Since $\sum_{d^{cens,k}} \sum_{\overline{d} \geq d^{cens,k}} p(d^{cens,k}, \overline{d}) = 1$, this implies that $\mathbb{E}[\mathbf{x}'\mathbf{G}\mathbf{x}] = \mathbf{A}$.

Similarly,

$$\text{(A.24)} \quad \mathbf{E_{xH}} = \sum_{d^{cens,k}} \sum_{\overline{d} \geq d^{cens,k}} p(d^{cens,k}, \overline{d}) \mathbb{E}[\mathbf{x}'\mathbf{H_k}\mathbf{x} | d^{cens,k}, \overline{d}]$$

$$\text{(A.25)} \quad = \sum_{d^{cens,k}} \sum_{\overline{d} \geq d^{cens,k}} p(d^{cens,k}, \overline{d}) \mathbb{E}[x_i (\frac{1}{d^{cens,k}} \sum_{v=1}^{d^{cens,k}} x'_{i(v)}) | d^{cens,k}, \overline{d}]$$

Next, given order irrelevance, for any $d^{cens,k}, \overline{d}$

$$\text{(A.26)} \quad \mathbb{E}[x_i (\frac{1}{d^{cens,k}} \sum_{v=1}^{d^{cens,k}} x'_{i(v)}) | d^{cens,k}, \overline{d}] = \frac{1}{d^{cens,k}} d^{cens,k} \mathbb{E}[x_i x'_{i(v)}] = \mathbf{A}$$

Since $\sum_{d^{cens,k}} \sum_{\overline{d} \geq d^{cens,k}} p(d^{cens,k}, \overline{d}) = 1$, this implies that $\mathbb{E}[\mathbf{x}'\mathbf{H_k}\mathbf{x}] = \mathbf{A}$.

Step 2: Expressions for $\mathbf{E_{HG}}$ and $\mathbf{E_{HH}}$

$$\text{(A.27)} \quad \mathbf{E_{HG}} = \mathbb{E}[\mathbf{x}'\mathbf{H}'_\mathbf{k}\mathbf{G}\mathbf{x}] = \sum_{d^{cens,k}} \sum_{\overline{d} \geq d^{cens,k}} p(d^{cens,k}, \overline{d}) \mathbb{E}[\mathbf{x}'\mathbf{H}'_\mathbf{k}\mathbf{G}\mathbf{x} | d^{cens,k}, \overline{d}]$$

Decompose as

$$\mathbf{E_{HG}} = \sum_{d^{cens,k}} \sum_{\overline{d} \geq d^{cens,k}} p(d^{cens,k}, \overline{d}) \mathbb{E}[(\frac{1}{d^{cens,k}} \sum_{v=1}^{d^{cens,k}} x_{i(v)})(\frac{1}{\overline{d}} \sum_{w=1}^{\overline{d}} x'_{i(w)}) | d^{cens,k}, \overline{d}] \tag{A.28}$$

$$= \sum_{d^{cens,k}} \sum_{\overline{d} \geq d^{cens,k}} \frac{p(d^{cens,k}, \overline{d})}{\overline{d} d^{cens,k}} \sum_{v=1}^{d^{cens,k}} \sum_{w=1}^{\overline{d}} \mathbb{E}[x_{i(v)} x'_{i(w)} | d^{cens,k}, \overline{d}] \tag{A.29}$$

$$= \sum_{d^{cens,k}} \sum_{\overline{d} \geq d^{cens,k}} \frac{p(d^{cens,k}, \overline{d})}{\overline{d} d^{cens,k}} \left(\sum_{v=1}^{d^{cens,k}} \mathbb{E}[(x_{i(v)} x'_{i(v)} | d^{cens,k}, \overline{d}] + \sum_{v=1}^{d^{cens,k}} \sum_{\substack{w=1 \\ w \neq v}}^{\overline{d}} \mathbb{E}[(x_{i(v)} x'_{i(w)} | d^{cens,k}, \overline{d}] \right) \tag{A.30}$$

Assumption 3 implies that this can be simplified as

$$\mathbf{E_{HG}} = \sum_{d^{cens,k}} \sum_{\overline{d} \geq d^{cens,k}} \frac{p(d^{cens,k}, \overline{d})}{\overline{d} d^{cens,k}} \left(\sum_{v=1}^{d^{cens,k}} \mathbf{B} + \sum_{v=1}^{d^{cens,k}} \sum_{\substack{w=1 \\ w \neq v}}^{\overline{d}} \mathbf{C} \right) \tag{A.31}$$

$$= \sum_{d^{cens,k}} \sum_{\overline{d} \geq d^{cens,k}} \frac{p(d^{cens,k}, \overline{d})}{\overline{d} d^{cens,k}} (d^{cens,k} \mathbf{B} + d^{cens,k} (\overline{d} - 1) \mathbf{C}) \tag{A.32}$$

$$= \sum_{d^{cens,k}} \sum_{\overline{d} \geq d^{cens,k}} \frac{p(d^{cens,k}, \overline{d})}{\overline{d}} (\mathbf{B} + (\overline{d} - 1) \mathbf{C}) \tag{A.33}$$

Reorder the summations and apply $p(d^{cens,k}, \overline{d}) = p(d^{cens,k} | \overline{d}) p(\overline{d})$. So,

$$\mathbf{E_{HG}} = \sum_{\overline{d}} \sum_{d^{cens,k} \leq \overline{d}} \frac{p(d^{cens,k} | \overline{d}) p(\overline{d})}{\overline{d}} (\mathbf{B} + (\overline{d} - 1) \mathbf{C}) \tag{A.34}$$

Since $\sum_{d^{cens,k} \leq \overline{d}} p(d^{cens,k} | \overline{d}) = 1$, this simplifies to

$$\mathbf{E_{HG}} = \sum_{\overline{d}} \frac{p(\overline{d})}{\overline{d}} (\mathbf{B} + (\overline{d} - 1) \mathbf{C}) \tag{A.35}$$

$$= \sum_{\overline{d}} p(\overline{d}) \mathbf{C} + \sum_{\overline{d}} \frac{p(\overline{d})}{\overline{d}} (\mathbf{B} - \mathbf{C}) \tag{A.36}$$

Noting that $\sum_{\overline{d}} p(\overline{d}) = 1$ and $\sum_{\overline{d}} \frac{p(\overline{d})}{\overline{d}} = \mathbb{E}[\frac{1}{\overline{d}}]$, this reduces to

$$\mathbf{E_{HG}} = \mathbf{C} + \mathbb{E}[\frac{1}{\overline{d}}] (\mathbf{B} - \mathbf{C}) \tag{A.37}$$

Similarly,

$$\mathbf{E_{HH}} = \mathbb{E}[\mathbf{x}' \mathbf{H}'_{\mathbf{k}} \mathbf{H_k} \mathbf{x}] = \sum_{d^{cens,k}} p(d^{cens,k}) \mathbb{E}[\mathbf{x}' \mathbf{H}'_{\mathbf{k}} \mathbf{H_k} \mathbf{x} | d^{cens,k}, \overline{d}] \tag{A.38}$$

Decompose as

$$\mathbf{E_{HH}} = \sum_{d^{cens,k}} p(d^{cens,k})\mathbb{E}[(\frac{1}{d^{cens,k}}\sum_{v=1}^{d^{cens,k}} x_{i(v)})(\frac{1}{d^{cens,k}}\sum_{w=1}^{d^{cens,k}} x'_{i(w)})|d^{cens,k},\overline{d}] \tag{A.39}$$

$$= \sum_{d^{cens,k}} p(d^{cens,k}) \left(\frac{1}{(d^{cens,k})^2}\sum_{v=1}^{d^{cens,k}} \mathbb{E}[x_{i(v)}x'_{i(v)}|d^{cens,k},\overline{d}] + \frac{1}{(d^{cens,k})^2}\sum_{v=1}^{d^{cens,k}}\sum_{\substack{w=1\\w\neq v}}^{d^{cens,k}} \mathbb{E}[x_{i(v)}x'_{i(w)}|d^{cens,k},\overline{d}] \right) \tag{A.40}$$

$$= \sum_{d^{cens,k}} p(d^{cens,k}) \left(\frac{1}{d^{cens,k}}\mathbf{B} + \frac{d^{cens,k}-1}{d^{cens,k}}\mathbf{C} \right) \tag{A.41}$$

$$= \sum_{d^{cens,k}} p(d^{cens,k})\mathbf{C} + \sum_{d^{cens,k}} \frac{p(d^{cens,k})}{d^{cens,k}}(\mathbf{B}-\mathbf{C}) \tag{A.42}$$

Since $\sum_{d^{cens,k}} p(d^{cens,k}) = 1$ and $\sum_{d^{cens,k}} \frac{p(d^{cens,k})}{d^{cens,k}} = \mathbb{E}[\frac{1}{d^{cens,k}}]$, this simplifies to

$$\mathbf{E_{HH}} = \mathbf{C} + \mathbb{E}[\frac{1}{d^{cens,k}}](\mathbf{B}-\mathbf{C}) \tag{A.43}$$

Combining (A.37) and (A.43),

$$\mathbf{E_{HG}} - \mathbf{E_{HH}} = \left(\mathbb{E}[\frac{1}{\overline{d}}] - \mathbb{E}[\frac{1}{d^{cens,k}}] \right)(\mathbf{B}-\mathbf{C}) \tag{A.44}$$

Step 3: Assemble pieces

Bringing it all together, note that

$$\mathbf{C_{HH}}^{-1}\mathbf{C_{HG}} = \mathbf{I} + \mathbf{C_{HH}}^{-1}(\mathbf{C_{HG}} - \mathbf{C_{HH}}) \tag{A.45}$$

$$= \mathbf{I} + \begin{bmatrix} \mathbf{E_{xx}} & \mathbf{E_{xH}} \\ \mathbf{E_{xH}}' & \mathbf{E_{HH}} \end{bmatrix}^{-1} \begin{bmatrix} 0 & \mathbf{E_{xG}} - \mathbf{E_{xH}} \\ 0 & \mathbf{E_{HG}} - \mathbf{E_{HH}} \end{bmatrix} \tag{A.46}$$

From Step 1, $\mathbf{E_{xG}} - \mathbf{E_{xH}} = 0$ and $\mathbf{E_{xH}} = \mathbf{A}$. Therefore,

$$\mathbf{C_{HH}}^{-1}\mathbf{C_{HG}} = \mathbf{I} + \begin{bmatrix} \mathbf{E_{xx}} & \mathbf{A} \\ \mathbf{A}' & \mathbf{E_{HH}} \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{E_{HG}} - \mathbf{E_{HH}} \end{bmatrix} \tag{A.47}$$

Define a matrix $\mathbf{D} = \mathbf{E_{HH}} - \mathbf{A}'\mathbf{E_{xx}}^{-1}\mathbf{A}$. Apply blockwise inversion to give

$$\begin{bmatrix} \mathbf{E_{xx}} & \mathbf{A} \\ \mathbf{A}' & \mathbf{E_{HH}} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{E_{xx}}^{-1} + \mathbf{E_{xx}}^{-1}\mathbf{A}\mathbf{D}^{-1}\mathbf{A}'\mathbf{E_{xx}}^{-1} & \mathbf{E_{xx}}^{-1}\mathbf{A}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{A}'\mathbf{E_{xx}}^{-1} & \mathbf{D}^{-1} \end{bmatrix} \tag{A.48}$$

Substitute (A.48) into Equation (A.47) and simplify:

$$\mathbf{C_{HH}}^{-1}\mathbf{C_{HG}} = \mathbf{I} + \begin{bmatrix} 0 & -\mathbf{E_{xx}}^{-1}\mathbf{A}\mathbf{D}^{-1}(\mathbf{E_{HG}} - \mathbf{E_{HH}}) \\ 0 & \mathbf{D}^{-1}(\mathbf{E_{HG}} - \mathbf{E_{HH}}) \end{bmatrix} \tag{A.49}$$

Substitute in the expression for $(\mathbf{E_{HG}} - \mathbf{E_{HH}})$ from Step 2, which yields

$$\mathbf{C_{HH}}^{-1}\mathbf{C_{HG}} = \mathbf{I} + \left(\mathbb{E}[\frac{1}{\overline{d}}] - \mathbb{E}[\frac{1}{d^{cens,k}}] \right) \begin{bmatrix} 0 & -\mathbf{E_{xx}}^{-1}\mathbf{A}\mathbf{D}^{-1}(\mathbf{B}-\mathbf{C}) \\ 0 & \mathbf{D}^{-1}(\mathbf{B}-\mathbf{C}) \end{bmatrix} \tag{A.50}$$

Since we have assumed $\beta_1 = 0$, Corollary 2 states that $\operatorname{plim} \hat{\alpha}^{cens,k} = \mathbf{C_{HH}}^{-1}\mathbf{C_{HG}}\alpha = \mathbf{C_{HH}}^{-1}\mathbf{C_{HG}} \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix}$.

Therefore,

$$\text{(A.51)} \quad \operatorname{plim} \hat{\alpha}^{cens,k} = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} + \left(\mathbb{E}[\frac{1}{\bar{d}}] - \mathbb{E}[\frac{1}{d^{cens,k}}] \right) \begin{bmatrix} -\mathbf{E_{xx}}^{-1}\mathbf{A} \\ \mathbf{I} \end{bmatrix} \mathbf{D}^{-1}(\mathbf{B}-\mathbf{C})\beta_3$$

Substitute back in the definition of $\mathbf{D}$ and rearrange, leaving the result

$$\text{(A.52)} \quad \operatorname{plim} \hat{\alpha}^{cens,k} = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} + \left(\mathbb{E}[\frac{1}{d^{cens,k}}] - (\mathbb{E}[\frac{1}{\bar{d}}] \right) \begin{bmatrix} -\mathbf{E_{xx}}^{-1}\mathbf{A} \\ \mathbf{I} \end{bmatrix} \left(\mathbf{E_{HH}} - \mathbf{A'E_{xx}}^{-1}\mathbf{A}\right)^{-1} (\mathbf{B}-\mathbf{C})\beta_3$$

APPENDIX B: DETAILS OF BIAS CORRECTION

I describe bias correction under two methods: (1) Full Method, and (2) Restricted Method. First, I discuss estimating parameters from the censored dataset, which are inputs into the Full Method only. Then I discuss estimation of $\mathbb{E}[\frac{1}{d^{true}}]$ from supplemental data, which is required for both methods. Finally, I bring these two parts together to construct the corrected estimator.

B.1. *Parameters Derived from Censored Data*

I first assume that the researcher has a censored dataset such that, for each agent, we observe

[1] Outcomes $y_i \in \mathbb{R}$

[2] Characteristics $x_i \in \mathbb{R}^m$

[3] Total number of ordered links $d_i^{cens,k}$

[4] For each $v \leq d^{cens,k}$, friends' characteristics $\bar{x}_{i(v)} \in \mathbb{R}^m$

From the censored data, we need consistent estimates of five parameters as inputs into $\hat{\mathbf{B}}_\mathbf{k}$. Table A.1 gives each of these parameters, a short description, and a suggested sample analogue estimator. Given the number of disjoint groups $S \to \infty$, these sample analogues are consistent.

TABLE A.1

PARAMETERS ESTIMATED FROM CENSORED DATA

Parameter	Description	Sample Analogue
$\mathbf{E_{xx}}$	variance of own characteristics	$\hat{\mathbf{E}}_\mathbf{xx} = \frac{1}{N}\sum_i x_i x_i'$
$\mathbf{A}$	covariance of own and links' characteristics	$\hat{\mathbf{A}} = \frac{1}{\sum_i d_i^k}\sum_i \sum_{z=1}^{d_i^k} x_i \bar{x}_{i(z)}'$
$\mathbf{B}$	variance of (individual) links' characteristics	$\hat{\mathbf{B}} = \frac{1}{\sum_i d_i^k}\sum_i \sum_{z=1}^{d_i^k} \bar{x}_{i(z)} \bar{x}_{i(z)}'$
$\mathbf{C}$	covariance of pairs of links' characteristics	$\hat{\mathbf{C}} = \frac{1}{\sum_i d_i^k (d_i^k - 1)}\sum_i \sum_{z=1}^{d_i^k} \sum_{\substack{w=1 \\ w \neq z}}^{d_i^k} \bar{x}_{i(z)} \bar{x}_{i(w)}'$
$\mathbb{E}[\frac{1}{d^{cens,k}}]$	average inverse number of censored links	$\hat{\mathbb{E}}[\frac{1}{d^{cens,k}}] = \frac{1}{N}\sum_i \frac{1}{d_i^{cens,k}}$

B.2. *Parameter Derived from Supplemental Data*

Supplemental data on the true degree distribution is needed to estimate $\mathbb{E}[\frac{1}{d^{true}}]$, the average inverse number of true links. In principle, any number of methods could be used to collect sufficient data. I describe two such methods here.

First, suppose that the researcher has access to Aggregate Relational Data (see, e.g., Breza et al., 2020; McCormick and Zheng, 2015), whereby a random sample of individuals is asked how many links each has in the relevant network. That is, we observe d_i^{true} for each of $i = 1, ..., N$ agents. From this we can construct a consistent estimate as

$$\hat{\mathbb{E}}[\frac{1}{d^{true}}] = \frac{1}{N}\sum_i \frac{1}{d_i^{true}} \tag{A.53}$$

Second, and alternatively, suppose that the researcher has access to data collected through two-step edge-wise sampling (see Conley and Udry, 2012). That is, within each group of size N_s: (1) a random subset of n_s individuals is sampled, and (2) each sampled individual is asked about the existence of a link to z_s other, randomly-chosen agents. Label f_{is} the number of links measured for agent i in school s with this procedure. For each individual i, conditional on (unobserved) degree d, f_{is} follows *hypergeometric distribution*. That is,

$$p(f_{is}|d) = \frac{\binom{d}{f_{is}}\binom{(N_s-1)-d}{z_s - f_{is}}}{\binom{(N_s-1)}{z_s}} \tag{A.54}$$

Therefore, for each agent, the likelihood of observing f_{is} is given as

$$p(f_{is}) = \sum_{d=f_{is}}^{N_s-1} p(f_{is}|d)p(d) \tag{A.55}$$

where $p(f_{is}|d)$ is given in Equation (A.54). Define $\mathbf{F}$ as a vector in which entry i corresponds to f_{is}. Now, we can write a likelihood function as

$$L(\mathbf{F}|p(1), p(2), ..., p(N_s - 1)) = \prod_i p(f_{is}) \tag{A.56}$$

which can be maximized by choosing $\hat{p}(1), \hat{p}(2), ..., \hat{p}(N_s - 1)$. Given these estimates, we then construct

$$\hat{\mathbb{E}}[\frac{1}{d^{true}}] = \sum_{d=1}^{N_s-1} \frac{1}{d}\hat{p}(d) \tag{A.57}$$

B.3. *Constructing Corrected* $\hat{\alpha}$

The next step is to construct $\hat{\mathbf{B}}_\mathbf{k}$. In the Full Method, $\hat{\mathbf{B}}_\mathbf{k}$ is constructed as a sample analogue of Equation (6) as given in Equation (A.58).

$$\hat{\mathbf{B}}_\mathbf{k} = \begin{bmatrix} \mathbf{I} & \hat{z}\hat{\mathbf{E}}_{\mathbf{xx}}^{-1}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-1}\left(\hat{\mathbf{B}} - \hat{\mathbf{C}}\right) \\ \mathbf{0} & \mathbf{I} - \hat{z}\hat{\mathbf{D}}^{-1}\left(\hat{\mathbf{B}} - \hat{\mathbf{C}}\right) \end{bmatrix}, \text{where } \hat{\mathbf{D}} = \hat{\mathbf{C}} + \hat{\mathbb{E}}[\frac{1}{d^{cens,k}}](\hat{\mathbf{B}} - \hat{\mathbf{C}}) - \hat{\mathbf{A}}'\hat{\mathbf{E}}_{\mathbf{xx}}^{-1}\hat{\mathbf{A}} \tag{A.58}$$

$$\hat{z} = \hat{\mathbb{E}}[\frac{1}{d^{cens,k}}] - \hat{\mathbb{E}}[\frac{1}{d^{true}}]$$

Alternatively, in the Restricted Method, which assumes $\mathbf{A} = \mathbf{C} = 0$, $\hat{\mathbf{B}}_\mathbf{k}$ can be constructed as in Equation (A.59), which is the sample analogue of Equation (7) in the text.

$$\hat{\mathbf{B}}_\mathbf{k} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \frac{\hat{\mathbb{E}}[\frac{1}{d^{true}}]}{\hat{\mathbb{E}}[\frac{1}{d^{cens,k}}]}\mathbf{I} \end{bmatrix} \tag{A.59}$$

Under either method, the final step is to combine $\hat{\mathbf{B}}_\mathbf{k}$ with $\hat{\alpha}^{cens,k}$, which can be estimated via regression from the censored data. The corrected estimate is then $\hat{\mathbf{B}}_\mathbf{k}^{-1}\hat{\alpha}^{cens,k}$ and its probability limit is α.

APPENDIX C: ORDERING

In the context of the linear-in-means model, when constructing an average of an individual's first k named links, all links are weighted equally. A worry about the linear-in-means models with ranked friendship data is that peer effects may be sensitive to ordering, such that the effect of the first friend is different than that of later-listed ones. Accordingly, sub-censoring may be less informative of patterns, as it drops links non-randomly. Such a concern may be particularly relevant with prompts, such as that in AddHealth, that ask respondents to list their closest friends first. Accordingly, to see if the results here are sensitive this the ordering prompt, I re-run both of the main analyses after reversing the order of links.

TABLE A.2

China Insurance Regression Results (Sub-Censored with Order Reversed)

Max Number of Links	1	2	3	4	5
Panel A: "OUT" Network					
$Treat_{is}$	0.029	0.028	0.027	0.028	0.030
	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)
$\overline{Treat}_{is}$	0.054	0.142***	0.211***	0.244***	0.291***
	(0.038)	(0.051)	(0.062)	(0.070)	(0.082)
R-squared	0.108	0.112	0.116	0.116	0.119
Panel B: "OR" Network					
$Treat_{is}$	0.028	0.027	0.027	0.028	0.031
	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)
$\overline{Treat}_{is}$	0.074	0.226***	0.268***	0.279**	0.311**
	(0.051)	(0.072)	(0.093)	(0.109)	(0.123)
R-squared	0.109	0.115	0.114	0.114	0.113

Notes: N = 1,255 in all specifications. All estimates correspond to specifications including village fixed effects and other controls as in Column (2) of Table 2 of Cai, de Janvry and Sadoulet (2015). Standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1. Order of friendship nominations in original data reversed.

For the China insurance data to show that the patterns in Table 1 are robust to reordering, I have reversed the original authors' ordering and re-run the sub-censoring analysis. Results are given in Table A.2, and these are consistent with the pattern we see in Table 1: estimates of the direct treatment effect are unaffected by censoring, while estimates of the peer effect are more attenuated the more censoring that occurs. Note that the results in Column 5 are identical in Tables 1 and A.2 since both use all friendship nominations. These results are unsurprising since the original dataset is essentially unordered.

For AddHealth, links are explicitly ordered within gender due to the survey prompt. That is, each respondent names their first female friend, second female friend, etc., and similarly for male friends. As a check of whether the results here are driven by this ordering, I re-estimate the main results with friendship nominations reversed. In doing so, I preserve each respondent's degree-by-gender. That is, if individual i named four female friends, I renumber friends 1, 2,3, and 4 to be friends number 4, 3, 2, and 1, respectively. I do the same for i's male friends.

These results for outcome GPA in All Subjects are in Table A.3. From this, we see that the results are qualitatively the same as those in the main text, in Table 2. That is, estimates on peer mean variables are attenuated with fewer friends observed, while those on one's own characteristics are biased in the opposite direction. These results are consistent with the analytic results under order irrelevance in Section 3 as well. Accordingly, this provides additional evidence that the patterns shown for the censored estimates are not driven by the ordered nature of the data.

TABLE A.3

ADDHEALTH REGRESSION RESULTS (SUB-CENSORED WITH ORDER REVERSED)

Censoring Rule (k)	1	2	3	4	5
Panel A: "OR" Network					
Age_{is}	-0.163***	-0.155***	-0.148***	-0.143***	-0.141***
	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)
$Grade_{is}$	0.169***	0.166***	0.157***	0.151***	0.150***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
$\overline{Age}_{is}$	-0.082***	-0.146***	-0.196***	-0.239***	-0.256***
	(0.010)	(0.016)	(0.019)	(0.022)	(0.025)
$\overline{Grade}_{is}$	0.079***	0.134***	0.187***	0.230***	0.244***
	(0.012)	(0.017)	(0.019)	(0.023)	(0.025)
R-squared	0.949	0.949	0.949	0.950	0.950
Panel B: "OUT" Network					
Age_{is}	-0.176***	-0.170***	-0.166***	-0.161***	-0.159***
	(0.013)	(0.013)	(0.012)	(0.012)	(0.012)
$Grade_{is}$	0.190***	0.191***	0.187***	0.182***	0.180***
	(0.014)	(0.014)	(0.014)	(0.015)	(0.015)
$\overline{Age}_{is}$	-0.072***	-0.112***	-0.156***	-0.194***	-0.216***
	(0.009)	(0.012)	(0.015)	(0.018)	(0.020)
$\overline{Grade}_{is}$	0.058***	0.088***	0.130***	0.167***	0.186***
	(0.010)	(0.014)	(0.016)	(0.019)	(0.021)
R-squared	0.950	0.951	0.951	0.951	0.951

Dependent Variable: GPA in All Subjects. N = 32,156 in all specifications in Panel A; N = 26,465 in all specifications in Panel B. Sample restricted to observations with non-missing data for all k. Standard errors in parentheses, clustered by school. School fixed effects included in all specifications. Coefficients for other covariates in Table A.5, Panel B not shown. *** p<0.01, ** p<0.05, * p<0.1. Order of friendship nominations in original data reversed.

APPENDIX D: SUPPLEMENTAL TABLES AND FIGURES

TABLE A.4

China Insurance Test Statistics

Comparison	1 vs. 5	2 vs. 5	3 vs. 5	4 vs. 5	All
Panel A: "OUT" Network					
$\alpha_1^{k,cens} = \alpha_1^{5,cens}$	0.150	0.047	0.712	0.129	2.072
	[0.698]	[0.828]	[0.399]	[0.720]	[0.723]
$\alpha_2^{k,cens} = \alpha_2^{5,cens}$	14.937	8.249	3.776	2.511	15.149
	[0.000]	[0.004]	[0.052]	[0.113]	[0.004]
$\alpha^{k,cens} = \alpha^{5,cens}$	15.020	8.449	4.960	2.950	15.887
	[0.001]	[0.015]	[0.084]	[0.229]	[0.044]
Panel B: "OR" Network					
$\alpha_1^{k,cens} = \alpha_1^{5,cens}$	1.740	1.348	1.259	1.330	2.767
	[0.187]	[0.246]	[0.262]	[0.249]	[0.598]
$\alpha_2^{k,cens} = \alpha_2^{5,cens}$	6.574	3.737	0.716	0.560	8.090
	[0.010]	[0.053]	[0.397]	[0.454]	[0.088]
$\alpha^{k,cens} = \alpha^{5,cens}$	6.631	4.329	1.748	1.698	9.761
	[0.036]	[0.115]	[0.417]	[0.428]	[0.282]

Notes: N = 1,255 in all specifications. All estimates correspond to specifications including village fixed effects and other controls as in Column (2) of Table 2 of Cai, de Janvry and Sadoulet (2015). Standard errors in parentheses, clustered by village. Table includes F-statistics of tests of equality across different levels of k. P-values in brackets, taken from $\chi^2(1)$ distribution in all but final column, and from $\chi^2(4)$ distribution in final column. Test statistics and p-values calculated from variance estimates that cluster by village.

TABLE A.5

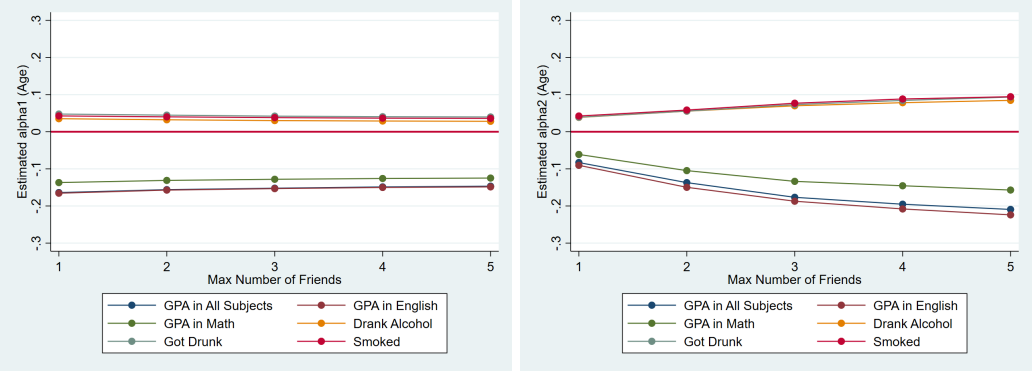
AddHealth Variables

	Min	Max	Mean	S.D.
Panel A: Outcomes				
Grade Point Average in All Subjects	1	4	2.891	0.782
Grade Point Average in English	1	4	2.857	0.979
Grade Point Average in Math	1	4	2.763	1.022
Has Drunk Alcohol (in last year)	0	1	0.560	0.496
Got Drunk (in last year)	0	1	0.317	0.465
Smoked (in last year)	0	1	0.359	0.480
Panel B: Independent Variables				
Female	0	1	0.509	0.500
Age	10	19	15.095	1.680
Grade	6	12	9.713	1.584
Hispanic	0	1	0.183	0.386
Black	0	1	0.177	0.382
Asian	0	1	0.068	0.252
Other Race	0	1	0.142	0.349
Born in the USA	0	1	0.903	0.297
Lives with Mother	0	1	0.925	0.264
Lives with Father	0	1	0.769	0.422

Notes: N = 70,364. Analysis dataset includes only students with friendship nominations and non-missing data for all variables in Panel B.

Figure A.1: AddHealth Estimated Coefficients (Sub-Censored, "OUT" Network)

(a) Coefficients for Age and $\overline{\text{Age}}$



(b) Coefficients for Grade and $\overline{\text{Grade}}$

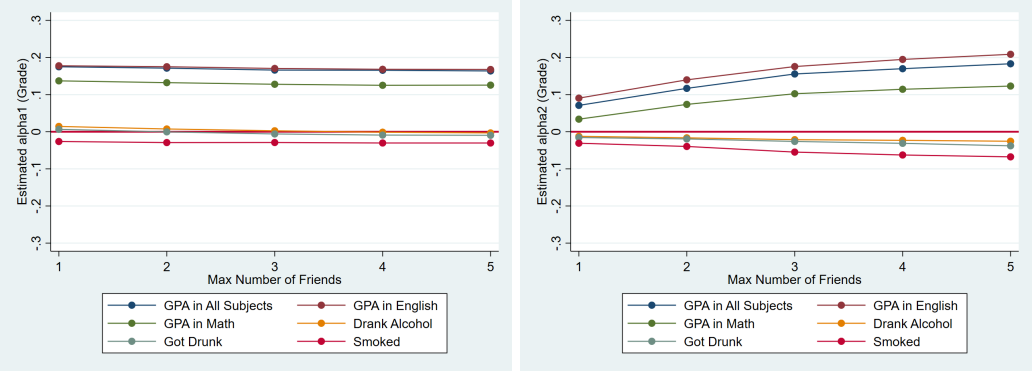


TABLE A.6

AddHealth Test Results ("OR" Networks)

k	1	2	3	4	All
Panel A: Test of $\alpha_1^{k,cens} = \alpha_1^{5,cens}$					
GPA in All Subjects	155.288	111.448	60.174	58.001	376.146
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
GPA in English	148.012	92.207	40.160	45.922	462.597
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
GPA in Math	96.211	72.390	51.194	25.795	219.962
	[0.000]	[0.000]	[0.000]	[0.004]	[0.000]
Drank Alcohol	165.825	140.620	51.995	25.018	329.786
	[0.000]	[0.000]	[0.000]	[0.005]	[0.000]
Got Drunk	244.504	197.025	81.792	28.866	456.434
	[0.000]	[0.000]	[0.000]	[0.001]	[0.000]
Smoked	183.509	173.837	81.852	27.715	353.391
	[0.000]	[0.000]	[0.000]	[0.002]	[0.000]
Panel B: Test of $\alpha_2^{k,cens} = \alpha_2^{5,cens}$					
GPA in All Subjects	209.609	175.563	116.766	103.404	576.122
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
GPA in English	214.341	126.959	80.299	88.179	546.253
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
GPA in Math	129.587	109.021	74.859	31.211	224.113
	[0.000]	[0.000]	[0.000]	0.001	[0.000]
Drank Alcohol	260.571	239.376	113.560	51.322	462.018
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Got Drunk	284.108	272.161	115.610	50.410	427.313
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Smoked	235.832	253.659	143.307	71.497	396.273
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Panel C: Test of $\alpha^{k,cens} = \alpha^{5,cens}$					
GPA in All Subjects	334.146	259.886	138.277	120.059	2053.200
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
GPA in English	284.733	165.510	101.718	111.758	2215.076
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
GPA in Math	191.158	148.207	87.859	50.116	932.435
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Drank Alcohol	329.328	290.660	148.470	97.244	915.475
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Got Drunk	378.576	305.978	172.390	98.168	1473.069
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Smoked	294.360	281.119	184.201	139.217	1069.634
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

Dependent Variable: GPA in All Subjects. N = 32,156 in all specifications in Panel A; N = 26,465 in all specifications in Panel B. Sample restricted to observations with non-missing data for all k. School fixed effects included in all specifications. Table presents test statistics for joint test across specifications as indicated. P-values in brackets, calculated from $\chi^2(10)$ distribution in Panels A and B for all but the final column, $\chi^2(40)$ distribution for the final column (10 is the number of exogenous regressors); $\chi^2(20)$ distribution in Panel C for all but the final column, $\chi^2(80)$ distribution for the final column. Test statistics and p-values calculated from variance estimates that cluster by village. All peer means constructed using "OR" network definition.

TABLE A.7

ADDHEALTH TEST RESULTS ("OUT" NETWORKS)

k	1	2	3	4	All
Panel A: Test of $\alpha_1^{k,cens} = \alpha_1^{5,cens}$					
GPA in All Subjects	165.631	94.586	60.058	33.530	286.461
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
GPA in English	104.154	60.437	30.212	22.965	184.220
	[0.000]	[0.000]	[0.001]	[0.011]	[0.000]
GPA in Math	106.532	53.270	21.976	13.108	197.287
	[0.000]	[0.000]	[0.015]	[0.218]	[0.000]
Drank Alcohol	132.672	91.802	41.043	29.268	200.493
	[0.000]	[0.000]	[0.000]	[0.001]	[0.000]
Got Drunk	202.312	126.883	62.254	34.362	329.187
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Smoked	123.085	88.781	47.337	21.023	222.057
	[0.000]	[0.000]	[0.000]	[0.021]	[0.000]
Panel B: Test of $\alpha_2^{k,cens} = \alpha_2^{5,cens}$					
GPA in All Subjects	228.195	180.909	90.568	52.478	462.902
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
GPA in English	154.653	100.037	51.890	27.555	372.817
	[0.000]	[0.000]	[0.000]	[0.002]	[0.000]
GPA in Math	136.317	77.681	26.884	15.293	267.923
	[0.000]	[0.000]	0.003	[0.122]	[0.000]
Drank Alcohol	169.146	124.570	61.728	48.255	311.082
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Got Drunk	242.175	188.773	95.116	51.719	424.439
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Smoked	168.086	155.775	82.169	43.753	330.671
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Panel C: Test of $\alpha^{k,cens} = \alpha^{5,cens}$					
GPA in All Subjects	307.021	199.191	103.309	65.963	1155.645
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
GPA in English	202.053	125.635	67.338	42.162	793.173
	[0.000]	[0.000]	[0.000]	[0.003]	[0.000]
GPA in Math	154.549	103.239	38.858	23.356	559.553
	[0.000]	[0.000]	[0.007]	[0.272]	[0.000]
Drank Alcohol	211.279	150.984	77.474	68.726	620.933
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Got Drunk	342.910	234.372	119.965	80.903	1457.497
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Smoked	209.190	209.027	112.003	102.246	977.712
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

Dependent Variable: GPA in All Subjects. N = 32,156 in all specifications in Panel A; N = 26,465 in all specifications in Panel B. Sample restricted to observations with non-missing data for all k. School fixed effects included in all specifications. Table presents test statistics for joint test across specifications as indicated. P-values in brackets, calculated from $\chi^2(10)$ distribution in Panels A and B for all but the final column, $\chi^2(40)$ distribution for the final column (10 is the number of exogenous regressors); $\chi^2(20)$ distribution in Panel C for all but the final column, $\chi^2(80)$ distribution for the final column. Test statistics and p-values calculated from variance estimates that cluster by village. All peer means constructed using "OUT" network definition.

TABLE A.8

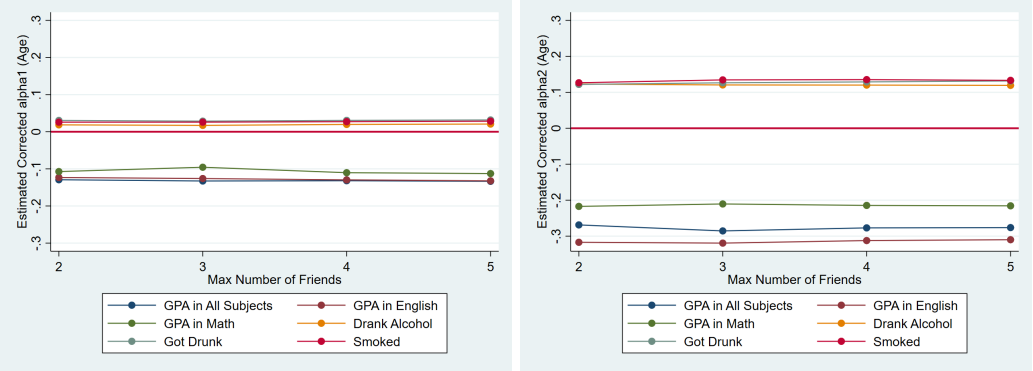
China Insurance Corrected Estimates ("OR" Network)

Max Number of Links	2	3	4	5
Panel A: Full Method				
$Treat_i$ ($\hat{\alpha}_1$)	0.031	0.035	0.037	0.034
	(0.033)	(0.033)	(0.033)	(0.033)
$\overline{Treat}_i$ ($\hat{\alpha}_2$)	0.302**	0.397***	0.325**	0.296**
	(0.150)	(0.151)	(0.133)	(0.126)
Panel B: Restricted Method				
$Treat_i$ ($\hat{\alpha}_1$)	0.031	0.034	0.037	0.034
	(0.033)	(0.033)	(0.033)	(0.033)
$\overline{Treat}_i$ ($\hat{\alpha}_2$)	0.331**	0.426***	0.334**	0.296**
	(0.165)	(0.162)	(0.136)	(0.126)

Notes: N = 1,255 in all specifications. All estimates correspond to specifications including village fixed effects and other controls as in Column (2) of Table 2 of Cai, de Janvry and Sadoulet (2015). "OR" network definition as used by original authors. Bootstrap standard errors in parentheses calculated from 1000 repetitions, sampled by village. *** p<0.01, ** p<0.05, * p<0.1.

Figure A.2: AddHealth Corrected Coefficients ("OUT" Networks)

(a) Coefficients for Age_{is} and $\overline{Age}_{is}$



(b) Coefficients for $Grade_{is}$ and $\overline{Grade}_{is}$