Three Essays on Econometrics of Networks

Jungyoun Kim

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Alan Griffith, Chair
Fahad Khalil
Xu Tan

Program Authorized to Offer Degree:
Economics

University of Washington

Abstract

Three Essays on Econometrics of Networks

Jungyoun Kim

Chair of the Supervisory Committee:
Alan Griffith
Department of Economics

This dissertation contributes three major contents to the econometrics of networks: application of discrete choice network formation models, identification of reduced form models with measurement errors in networks, and finding efficient sampling methods of epidemics over networks.

In the first chapter, we investigate the impact of the initial academic social network, formed from advisor-advisee relationships and coauthorships, for economics Ph.D. students (advisees) in the U.S. on their early stage productivity. We define the *academic social network* as a union of i) an advisor-advisee network and ii) a coauthorship network. We model the advisor-advisee relationships with a preferential attachment-like process based on a discrete choice model and find that advisees show weak gender homophilic preferences when choosing advisors. We further model early stage coauthorship formation of advisees through a bipartite network setup, also based on a discrete choice model, and find that advisees prefer to choose projects that are coauthored with their advisors during their graduate studies. Given the *academic social network* through the two networks, we find that the corresponding network statistics for advisees have significant positive correlation with early stage output but find weak evidence of gender causing difference. Through simulated synthetic data, we show that for advisees, in average, preference based decision making leads to individual level percentage-wise productivity gain but loss in the aggregate level, compared to random matching to advisors and projects. This implies that a preference based allocation of advisors to advisees is less efficient in the social planner's view.

In the second chapter, we consider the effects of the mismeasurement of networks on reduced-form peer-effect linear-in-means and linear-in-sums estimates. Applied researchers frequently estimate network-based peer effects models using observed network data that includes only a subset of the true links. Our results require an assumption that the expected covariance of characteristics between linked agents is the same regardless of whether the link is observed or not. Analytic results show that the linear-in-means peer effects estimate is in general attenuated, and this is a special case of "classical" measurement error. In contrast, linear-in-sums direct and peer effect estimates may be attenuated, augmented, or consistent; the inconsistency depends upon the missingness mechanism and the relationship between the network and covariates. We demonstrate the effect of mismeasured links in both models using two datasets and through simulations. These results show that the effects of mismeasured networks on subsequent estimands is quite sensitive to the parameter that is being estimated.

In the last chapter, we present a feasible version of the Neyman allocation for sampling epidemics over networks. Our method requires knowledge of only the first moments of the degree-based strata and an epidemic model that captures the dynamics of the diffusion process. Through simulations on randomly generated networks, we demonstrate that the optimal Neyman allocation and our proposed methods show efficiency gains over simple random sampling, particularly in the early stages of an epidemic. The feasible method closely approximates the performance of the optimal method while being implementable in practice. Our findings can inform sampling strategies for monitoring real-world epidemics given limited resources.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor and chair, Alan Griffith, for not only his research advice and guidance but also the encouragement and friendship that pulled me through this journey. I would also like to thank my committee members Xu Tan, Fahad Khalil, and Miruna Buta for their attention and support. A special thanks goes to CFRM director Tim Leung for numerous teaching opportunities and Jason Bigenho, my internship manager, for the great experience and mentorship.

I also thank my family and friends for their support during this challenging journey. Thank you to my wife Susie and son Eusheen for their love, sacrifices, and having faith in me. Thank you to my mom, dad, and sister for their unconditional love and support. Lastly, thank you to Dadmehr Didgar and my fellow UW Econ graduate students for creating a supportive environment that made this journey truly enriching.

DEDICATION

To Susie, Eusheen, Mom, and Dad.

Chapter 1

CHOOSE YOUR ADVISER WISELY: ENDOGENOUS ADVISOR-ADVISEE RELATIONSHIP AND EARLY STAGE COAUTHORSHIP ON RESEARCH OUTPUT

In this chapter, we investigate the impact of initial *academic social networks* on early-stage productivity for economics Ph.D. students in the U.S. We model advisor-advisee relationships and coauthorship formation using discrete choice models, finding weak gender homophily in advisor selection and a preference for students to coauthor with advisors. Corresponding network statistics show significant positive correlation with early-stage output. Through simulations, we demonstrate that preference-based network formation results in individual-level productivity gains but lower aggregate output compared to random matching and formation. This suggests advisee preference-based allocation of advisors is less efficient from a social planner's perspective. Our novel study provides insights on the effects of the formation stage of advisor and coauthor relationships on research productivity.

1.1 Introduction

Perhaps the most important first decision a graduate student makes during their program is choosing their advisor. That is, advisors are the closest sources of information, guidance, networking, and collaboration, which makes them to be the most important asset for a student to acquire before graduation.[1] Numerous work has been done in emphasizing the importance of an advisor to their students, both qualitatively and quantitatively, but there has been no attempt on measuring the effect of the allocation process of said asset. This paper takes a novel approach by modeling

[1]For economics; Hilmer and Hilmer (2009),García-Suaza, Otero and Winkelmann (2020). For other disciplines and academia in general; Artiles and Matusovich (2022), Litalien and Guay (2015), Lovitts (2001), Sauermann and Roach (2012). For qualitative work, see Zhao, Golde and McCormick (2007), Dericks et al. (2019), Barnes and Austin (2009)

the advisor-advisee formation process of students in economics Ph.D. programs based in the U.S. through a network formation model. Then we study the effects of those connections on early stage coauthorship formation, and further see how different allocation processes of advisors effect the students at the individual level and the aggregate level.

This study starts with the question of "what if a student met a different advisor?" In order to address this, we first start with the obvious decision process: a student choosing an advisor. During their first couple years in the program, a typical student in a economics Ph.D. program in a U.S. based institute chooses their advisor from a pool of choices based on their preference.[2] This forms what is called a genealogy network, namely, and advisor-advisee network. Then, given the relationship, a student starts their early stage research, under some influence by their advisor, if not in a collaborative effort.[3]. If a student engages in a collaboration, then they form a coauthorship network with their coauthor. Lastly, the networks the students formed and the early stage research would further lead to more research output as their academic career expands. Thus, we can answer the question through this channel by modeling each step.

This brings us to our first contribution. We take a novel approach of the advisor-advisee network formation process using the genealogy tree data of the economics literature community members presented by the IDEAS RePEc initiative. Despite advisors playing a key role for the career of an academic as shown above, the literature lacks a quantitive approach on how the process of choosing one works. The two main things to consider in modeling this process is i) advisees have a pool of advisor to choose from (oppotunity set) and ii) the network shows a preferential attachment behavior – i.e., advisors with more students are likely to be more attractive. In order to incorporate these two points, we employ a discrete choice based preferential attachment-like model to formulate the growth process of the network. Namely, we use the number of past students and pairwise attributes as variables in a restricted-set multinomial conditional logit model. The restriction here refers to the difference in the advisor opportunity sets of which the students can choose from. We find that the network indeed resembles a preferential attachment process, but also have weak

[2]We assume a student knows if they will be rejected so there exists a partial equilibrium.

[3]Some students may have independent research before meeting their advisor so we consider multiple projects.

gender homophily.[4]

Next, we model the coauthorship formation process. Similar to how students chose their advisor, students decide to participate in research projects from a pool of perspective projects and for those who choose a collaborative work get to form a coauthorship network with the coauthors. Following Hsieh et al. (2022), we model this coauthorship relationship as a bipartite network but distinct ourselves by modeling the formation process with a discrete choice model. As the genealogy network formation, the restricted-set multinomial conditional logit model allows restricted choice sets and under a choice independence assumption, allows for multiple choices as well. We find that, while students are likely to work on a single author project, if they do collaborate, they are likely to work with their advisor or their advisor's coauthor, but not other faculty members (potential advisors to be exact). We also find that with this controlled, we don't observe gender homophily in the early stage coauthorship network.

The last stage allows us to form out second contribution, which is quantifying the difference in allocation of advisors among students – i.e. answering the initial question. In order to do so, we construct a production function that projects the networks statistics of the union network of the two on to a output measurement. Then, we generate synthetic data from the genealogy network formation model, i.e. match different advisors to students. Next, we generate the coauthorship network to match different projects to students, which would be conditional on who their new advisor became. Then using the production function, we calculate the output for each new case.[5] This counterfactual study would allow use to compare the output depending on how the advisors and project are matched. We find that when they are matched according to the model, there are positive percentage gains in average at individual levels compared to a case when advisors and projects are matched randomly. However, we find that it is the opposite in the aggregate level, the average total output is greater for random matching. This finding implicates that, decisions made on preferences are dominant strategies for individuals but not an efficient state in the views of a

[4]Gender information is found through *Gender-API.com*. Details are in section 1.3.1.

[5]These synthetic data generation processes require some strict assumptions. Details are in section 1.2.4 and section 1.3.2.

social planner.

Our work shares mainly three branches of the current literature. The first is the area of network formation, a prominent and widely expanding area.[6] We specifically align with the works of Wichmann, Chen and Adamowicz (2016), Overgoor, Benson and Ugander (2020), and Gupta and Porter (2022) which employ a multinomial logit model as a network formation process. Especially for the genealogy network formation, we follow Overgoor, Benson and Ugander (2020) where the connection of discrete choice modeling and preferential attachment is described. Our view on defining the coauthorship network as a bipartite network resembles the work of Hsieh et al. (2022), but the discrete choice modeling of the formation process follows the works of Fu et al. (2017) and Yeung (2019).

The second area is related with coauthorship and collaborative research. Related literature in the field of economics date back to Sauer (1988), but the works of Goyal, van der Leij and Moraga-González (2006) and Azoulay, Zivin and Wang (2010) extend the concept to the network topology, while Fafchamps, Goyal and van der Leij (2010) further studies the formation process of coauthorship networks. Our modeling of the production function borrows the idea of Ductor et al. (2014) in which they find that coauthorship network statistics is useful in predicting future output of a researcher. More recent studies of Ductor, Goyal and Prummer (2021) find how different network characteristics by gender explain the output inequality in research.

The third area are quantified empirical studies related to the advisor-advisee relationship in the discipline of economics. Our work is directly related to the those of García-Suaza, Otero and Winkelmann (2020) and Hilmer and Hilmer (2009), where they find how the quality of the advisors and institutions are positively correlated with the students' early stage performance, especially how students coauthoring with their advisors outperform others, using different datasets.[7] However, we extend the work further to allowing the network related exogenous variables to be endogenous and

[6]Graham (2015), Chandrasekhar (2016), de Paula (2017) and de Paula (2020), provide broad reviews in econometrics of network formation.

[7]Some other work in the area, though not directly related, include: Tol (2021) for advisor-student relationship and noble laureates, Colussi (2018) and Brogaard, Engelberg and Parsons (2014) for advisor-advisee connection and publication.

formulate the process for counterfactual studies.

While our scope of research does not extensively study homophilic preferences, it is well known that social networks exhibit gender and racial homophily.[8] We choose to use these as control variables in all models thus as a bi-product we observe the strength of homophilc preference in the network formation models as well as gender inequality in research output. Comparable work related to our findings are the likes of Hilmer and Hilmer (2007), Gaule and Piacentini (2018), and Pezzoni et al. (2016), which study the impact of advisor-advisee gender matches on research output in economics and other disciplines. We find mixed results compared to the previous literature.

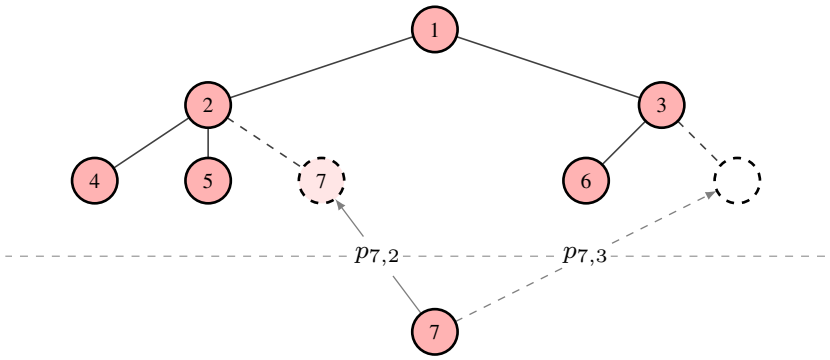
The remaining of the chapter has the following structure. We introduce the methodology of the study in section 1.2, describe the data collection process and definitions of key variables in section 1.3, report the empirical findings in section 1.4, and share remarks and conclude in section 1.5.

1.2 Methodology

In this section, we first introduce the networks we use in our analysis – the genealogy network and the coauthorship network – and it's corresponding growth processes. Then we define an *academic social network* by combining these two networks which can be seen as human capital a student can accumulate during their studies in graduate school. Given such, we show our empirical strategy on how to measure the impact of the formation processes through a production function of the early stage research of those students. The concepts here are illustrated in a simple manner.[9]

Before moving on, we clarify some terminology. The expression *student* and *advisee* is used interchangeably henceforth. Each individual is an *author*, who can have a label of either advisee and/or advisor or neither. Each paper or working paper, published in a journal or working paper series respectfully, would be addressed as a *project*.

Figure 1.1: Genealogy Network Growth Process – Example



Note: New advisee node 7 faces a pool of potential advisors node 2 or node 3 and decides to choose node 2 with the probability $p_{7,2}$. Since there are only two choices possible, $p_{7,3} = 1 - p_{7,2}$.

1.2.1 Genealogy Network and Growth

The genealogy network describes the advisor-advisee relationship. By nature, it is a tree network.[10] The upper part of Figure 1.1 illustrates an example of a genealogy network. We can see that each node – an author – could be either and advisor (node 1) or an advisee (nodes 4, 5, and 6) or both (nodes 2 and 3). The lower part of Figure 1.1 consists a potential advisee – node 7 – who selects node 2 as their advisor based on a preference structure, from a pool of node 2 and 3 as potential advisors (we assume node 1 is not available here). This preference structure would determine the selection probability for each node, denoted as $p_{7,2}$ and $p_{7,3}$ respectfully in the figure.

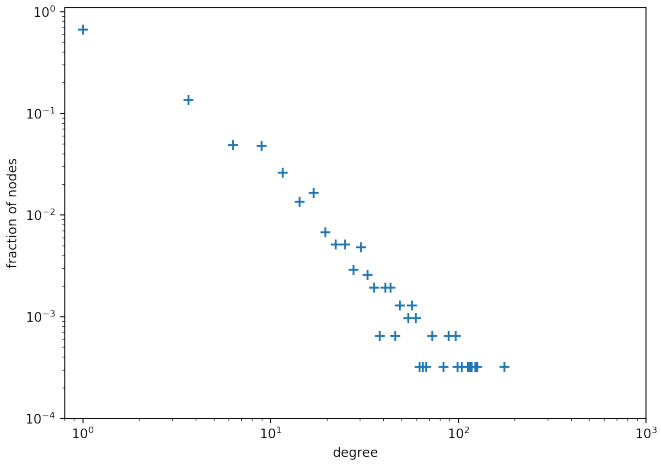
Obtaining these probabilities allows us to formulate the growth process of this network and thus understand how advisee-advisor relationships are formed. We assume that each advisee faces a pool of potential advisors (node 2 and 3 in the example) and would choose an advisor conditional on advisor specific (e.g. gender) and pairwise attributes (e.g. gender match). Formally, we define

[8]McPherson, Smith-Lovin and Cook (2001)

[9]Detailed mathematical definitions are in Appendix A.1.

[10]In our dataset, some advisees have two advisors but since the data indicates who is the first and second, we discard the second advisor for our analyses; the number of those how had two advisors was less than 0.5% of the total sample.

Figure 1.2: Log-Log plot of the unweighted out degree distribution of the genealogy network



Note: The linear slope of the data points suggest that the degree follows a power law / pareto distribution. This is of the out degree of the network since the in degree is always 0 or 1.

the probability of advisee i choosing advisor j with and restricted-set conditional logit model as

$$p_{i,j} = P(advisee_i = j | j \in AdvisorPool_i) = \frac{\exp(\alpha d_j + z'_{ij}\delta)}{\sum_{k \in AdvisorPool_i} \exp(\alpha d_k + z'_{ik}\delta)}$$

where variable d_j denotes the number of students advisor j has at the time of the selection and z_{ij} denotes a vector of pairwise attribute variables. In our example in Figure 1.1, we have $AdvisorPool_i = \{2, 3\}$ and thus $d_2 = 2$ and $d_3 = 1$ as the advisor specific attribute. For the pairwise attributes, in case of categorical information such as gender, if node 2 advisor and node 7 advisee are both males, then $z'_{27} = (1, 0)$.[11]

This model is an augmented form of a "preferential attachment with fitness" process as de-

[11]For categorical data, the dimension of vector z'_{ij} is the number of all categories. For example, if we only have gender data, the dimension would be 2, where each element index is the category for the advisee gender type and the elements are corresponding dummy variables that take value 1 if the gender are the same. So, for a female-female advisor-advisee match, $z'_{ij} = (0, 1)$, and male-female or female-male, $z'_{ij} = (0, 0)$.

scribed in Overgoor, Benson and Ugander (2020) where we use the number of past students instead of the degree of each advisor and have an restricted choice set setup.[12] Our assumption on this approach is based on the nature of the genealogy tree where there are multiple advisee authors connected to one advisor author. Also, the fact that it is more common, at least in the field of economics, for a student to propose to a professor of their choice after observing their characteristics.[13] We also consider the fact that each advisee student has a limited number of advisors to choose from, constrained by both time and place. Figure 1.2 illustrates the out degree of the genealogy network where the downward sloping linear trend further supports the usage of this approach.[14]

1.2.2 Early Stage Coauthorship Network Formation

While a coauthorship is a relationship between two authors, the component that connects the authors is a project that they participate in. Thus, to form a coauthorship network, connection, an author should be choosing a project, conditional on the information of potential coauthors. Formally, we form the coauthorship network through a author-to-project bipartite network as in Hsieh et al. (2022). Figures 1.3 and 1.4 illustrate the process.

While a coauthorship is a relationship between two authors, the component that connects the authors is a project that they participate in. Thus, to form a coauthorship network, connection, an author should be choosing a project, conditional on the information of potential coauthors. Formally, we form the coauthorship network through a author-to-project bipartite network as in Hsieh et al. (2022). Figures 1.3 and 1.4 illustrate the process.

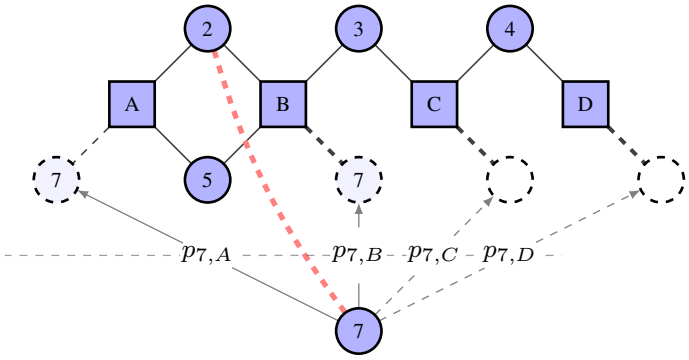
In the upper part of Figure 1.3, we have a bipartite network with 4 authors (nodes 2, 3, 4, and 5) and 4 projects (nodes A, B, C, and D). The projection of this network on to the set of authors would yield the coauthorship network in the upper part of Figure 1.4, where the width (weight) of

[12] A preferential attachment model has probability of $p_{i,j} = \frac{d_j^\alpha}{\sum_k d_k^\alpha}$.

[13] We assume that the advisor author – student – has enough information that they know whether their proposal will be rejected or accepted.

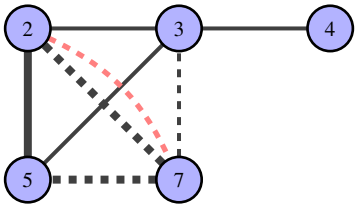
[14] The degree distribution of a network built from a preferential attachment process will have a pareto distriubtion, thus having a downward sloping linear trend of the log-log plot.

Figure 1.3: Project Selection Process – Example



Note: Advisee Node 7 is the potential node to be joining the network above. The red dashed line indicates the advisor-advisee relationship between node 7 and 2. With each corresponding probability, node 7 can choose among a pool of projects; A, B, C, and D.

Figure 1.4: Coauthorship Formation Process – Example continued



Note: Given the choice of node 7, project A and B, projecting the bipartite network on the author set results in the coauthorship network shown above. We can see that the difference in the line width represents the difference in numbers of projects done between coauthors.

the edges are proportionate to the number of projects two authors share. Since author 2 and 5 share two project A and B, the edge connecting the two are thicker (has twice the weight compared to other edges).

Given this configuration, the coauthorship network growth process starts with the advisee author node 7. Recall that advisee node 7 formed an advisor-advisee relationship with author node 2 from the example in Figure 1.1, which is denoted as the dashed red lines. Conditional on this advisor-advisee relationship and a preference structure, advisee node 7 chooses project A and B with the corresponding probabilities $p_{7,A}$ and $p_{7,B}$ over the set of candidate projects A, B, C, and D. Projecting this network on the set of authors results in the coauthorship network in Figure 1.4 where we see how the new edges – blacked dashed lines – from advisee node 7 to author nodes 2 and 5 are thicker (twice the weight) than that of the connection to node 3 since both projects A and B involves authors 2 and 5 while author 3 participates in only project B. Thus, if we can formulate the preference structure for the decision process of advisee node 9 choosing projects, we can model the growth process of the coauthorship network.

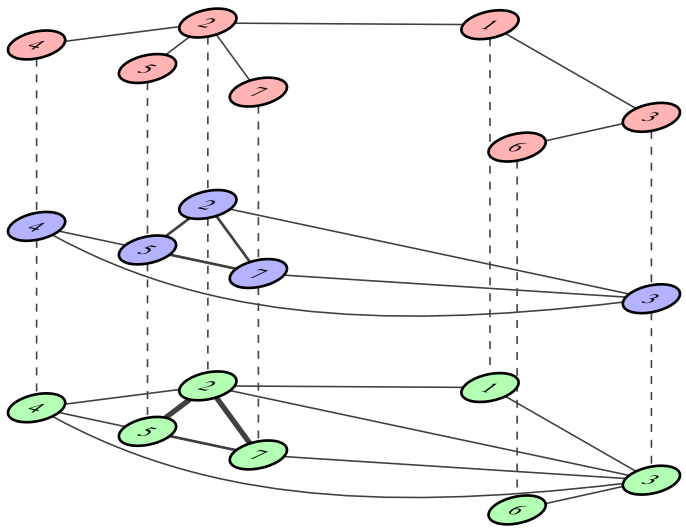
We model the preference structure similar to that of the genealogy network growth process. Formally, with the assumption that each decision is independent, we define the probability of author i choosing project s_n as

$$p_{i,s_n} = P\left(advisee_i = s_n | s_n \in ProjectPool_i\right) = \frac{\exp(q'_{is_n}\theta)}{\sum\limits_{k \in ProjectPool_i} \exp(q'_{ik}\theta)}$$

where vector q'_{is_n} is the vector of pairwise attributes between author i and the coauthors of project s_n.[15] For example, one of the variables is a dummy variable that indicates whether a coauthor of

[15] s_n for $n \in \{1, ..., capacity_i\}$ is the n'th project with a maximum value that takes is the working capacity of advisee i, i.e. the total number of projects advisee i participated in. For example, if advisee i worked on three projects during their graduate studies that $capacity_i = 3$. In the works of Gupta and Porter (2022), they assume the independence of choices in cases with mulitple choices in a discrete choice model based network formation. In their setup, the parameters are allowed to vary for each individual (heterogeneous preferences) due to variation introduced by multiple choices. Thus the probability for each individual's decision (of mulitple choices) is the product of the likelihood for each choice. In contrast, we assume homogeneous preferences due to some authors only having one choice observed and thus have a simple likelihood function as here.

Figure 1.5: Academic Social Network – Example



Note: The first and second layer is the genealogy network and coauthorship network respectively. The third layer is the academic social network which can be seen as a union projection of the two networks to the last layer. The edge width does not reflect the depreciation, just the additiveness.

project s is the advisor of author i. In our example above, $q_{7,B} = 1$ and $q_{7,A} = 1$ but $q_{7,C} = 0$ since the coauthor of project A and B includes author node 2, who is author node 7's advisor, but not in project C. Details on how we construct the variables are described in section 1.3.2 in detail.

1.2.3 Academic Social Network

We define the *academic social network* using the union of the two social networks; the genealogy and early stage couathorship network. This network can be seen as human capital a student can form during their studies in graduate school to use it for future production: research.

Formally, by having a same additive measurement for the edges of each network, we can simply add one network on top of the other. In order to have the same measurements on each network, we weight each edge by the inverse of the years that have past since the event connecting the two nodes happened. That is, for the genealogy network, the weights would be the inverse of years

after graduation, and for the coauthorship network, the weights would be the inverse of the years after publication to a journal or a posting of a working paper. We choose this weighting scheme in the view of considering each publication or the advisor-advisee relationship as an *academic social encounter*, which depreciates over time. It is natural to think that each connection to be less stronger as time goes by, even for advisee-advisor relationships. That is, the initial advisor-advisee connetion has much weight in fresh graduates' academic network, but would gradually decrease unless an advisee frequently cooperate with their advisor throughout their career.

Figure 1.5 illustrates the concept. Continuing on the previous examples, the new advisee node 7 participates in project A and B conditional on the fact that they chose node 2 as their advisor. The upper and middle layers illustrate this new state, the new genealogy network and new coauthorhip network respectively. Then, we combine the two networks by taking the union of the nodes, all nodes of 1,..., 7, while adding the weights of the corresponding edges.[16] The network in the last bottom layer is the result, which is the academic social network advisee node 7 would be in, by the time of their graduation.

1.2.4 Empirical Strategy

We aim to measure the impact of the academic social network and its formation process on an advisee's early career performance. To measure the impact, we do so by constructing a log-linear production function as

$$\bar{y}_i = \exp\left(w_i'\gamma + x_i'\beta + \epsilon_i\right)$$

where $\bar{y}_i$ is the output measure as we define in section 1.3.2.

In this function, the parameter of interest is γ which measures the impact of the academic social network as w_i is a vector of the network statistics (human capital). Vector x_i which denotes the control variables, namely, the fixed effects of each advisee such as institution, gender, and region of origin.

Next, to measure the impact of the two network formation processes, we conduct a counterfac-

[16]Formal definition is provided in Appendix A.1.1.

tual study. Specifically, we use synthetic data generated by each process to obtain the output under alternative circumstances – different advisor and thus different coauthors – and compare the values to that of what the model predicts with the predicted output from the original data. The synthetic data generation process starts by constructing a genealogy network based on the formation model. For each advisee, we random sample an advisor-author proportionately to the predicted probabilities assigned to each candidate in the pool of advisors. This allows us to construct a new genealogy network, where the original connections are removed and replaced by the newly generated edges. We assume that the graduation year of the advisee does not change, i.e. not dependent on the advisor-author, so the weights on the new edges are calculated accordingly.

Given the new genealogy network, we predict new probabilities using the fitted coauthorship network formation model. In this process, we impose three assumptions that allows us to generate plausible data. The assumptions are as follows.

[1] *The projects and its original authors, sans the advisee, are fixed*. That is, all candidate projects are, in some way, meant-to-happen regardless of who the newly joining coauthor would be which could be seen as a rather strong assumption. However, given the fact that advisees are newcomers to academia while the original authors are mostly likely to be experienced enough to have their on going research pipeline, the formation of the coauthorship could be seen more like a "joining as a branch" from the advisee's prospective rather than a "starting a whole new different project."

[2] *The capacity of each project is fixed*. The number of newly generated joining authors for each project should be the same as the original number of authors. For example, if there were three original authors, two of which are not part of the advisee pool, then only one new advisee can join the project. Similarly, if only one of the original author is not from the advisee pool, then two advisees can jointly join the project. This assumption prevents a project from overwhelming with newly joining advisees.

[3] *The capacity or ability for each advisee is fixed*. As denoted by r_i in section 1.2.2, the

number of projects an advisee joined during their first stage of their career is fixed. This includes the number of solo projects. This assumption also maintains the average degree of projects to be at a realistic level.

With the assumptions above, for each advisee and their pool of projects, we random sample r_i number of projects without replacement proportionate to the predicted probabilities from the fitted model. During this process, per assumption 2, each advisee selects from an exhaustive pool of projects on a first-come-first-serve basis. If a chosen project has already been taken, then we draw the next random sample with the second highest weight. The order of choosing is shuffled for each iteration of data generation to avoid matching bias.

After collecting all new author-project pairs and constructing a new coauthorship network, we construct the adacemic social network and obtain the corresponding network statistics for each advisee. Then, with the original fixed effects of each individual, we finally collect the corresponding output through the fitted production function. By comparing the output distribution, we gauge the effect of the network formation processes.

1.3 Data and Variables

1.3.1 Data Collection

The data is constructed from two sources; the RePEc initiative and *Gender-API.com*.[17] The former assembles a bibiliographic meta database from over 2000 providers relevant to economics including all major publishers and research outlets whereas the latter is an AI powered search which provides services on determining gender and country of origin by name. We collect the necessary data to identify the network structure between the authors and use the names of each author to find the corresponding gender and region of origin. We also construct the output variable based on the publication records for each author.

[17] *Gender-API.com* is an online platform that estimates a gender and region of origin based on the first or last name (or both), email, and IP address using AI and machine learning models. Their data sources are publicly available data, governmental data and manual additions/corrections. Ductor, Goyal and Prummer (2021) uses this service to construct their data set, which is used in identifying the productivity difference between male and female authors.

Starting with the RePEc Genealogy project database, we collect the information of advisees, advisees' advisor, year of graduation, and the institution they graduated from Then, we use the RePEc Author Service which contains information of each author's name and project – published journal or posted working paper – record. Cross-referencing this with the RePEc Publisher data, which includes a list of authors' names and the year of publication/posting for each project, allows us to construct the time varying coauthorship network.

We collect a total of 59,069 authors and 680,461 projects for the time varying coauthorship network, out of which there are 10,597 authors who are connected in the genealogy network as well. Given the base dataset, we apply a series of filters to select a pool advisees to fit the two network growth models and the production function. First, we select those who graduated between 2006 to 2015 from the department of economics of the top 25% US based institutions as ranked by IDEAS RePEc.[18] Next, we collect those who have at least one effective publication or working paper – project that has a positive output measure – throughout the five years after their graduation and also at least one publication or working paper posted during their graduate studies , i.e. 3 years before graduation or 1 year after to be exact. Finally, we remove the advisees with only one choice in their pool of advisors or pool of projects for each network growth model to ensure identification.

Given the set of the remaining advisees, we identify i) the advisor-authors included in the pool of advisors along with ii) the authors of the projects each advisee can choose for each network growth model, then use *Gender-API.com* to collect the gender and region of origin information.[19] Authors without any gender nor region of origin information are removed from the pool and the filters are applied accordingly. This leaves us with a total of 431 advisees for a pool of 457 advisor-authors and 2,203 projects.

We use the RePEc Publisher database to construct the time varying output variable by collecting the journal and working paper series information. The database includes over 4000 journals and 6000 working paper series, which we select a subsample of 1000 journals and working paper series

[18]Ranking as of Sep. 2023 based on all authors and all publication years.

[19]By providing the full name of an author, the API returns a binary result of gender with a probability and a list of possible region of origins with a binary probability of each name coming from each region.

based on the ranking in CitEc, a RePEc service that provides citation analysis. More detail on how we construct the output variable is described in section 1.3.2.

1.3.2 Variable Descriptions

Output Measure

We define the time varying output as the research output measure from Ductor, Goyal and Prummer (2021). That is, the sum of the number of publications for the past five years weighted by the quality of the journal or working paper series and discounted by the number of coauthors, for each year. Formally,

$$Y_{it} = \sum_{s=1}^{S_{i,t}} \frac{AIS_s}{(\text{no. of authors})_s}$$

where $S_{i,t}$ is the set of all projects author i published or posted in a working paper series from time t to $t-4$ and AIS_s is the a *article influence score* of the journal or working paper series of project s, which is a measure of quality for said journal or working paper series.

Following Bergstrom, West and Wiseman (2008), we calculate the AIS for journal or working paper series j at time t as

$$AIS_{jt} = \frac{EF_{jt}}{a_{jt}}$$

where EF_{jt} is the *eigenfactor* of journal or working paper series j at year t which solves the following recursive problem

$$EF_{jt} = \sum_{k \in \mathcal{K}} \frac{c_{jk,t}}{\sum_k c_{jk,t}} EF_{kt},$$

and a_{jt} is the normalized project share vector, where each element is the number of all projects in journal or working paper series j divided by the total number of projects in the same sample window collected for time t. Variable $c_{jk,t}$ is the jk-th element in the citation matrix – a 1000 by 1000 matrix given the data set – where each element is the total number of projects in journal or working paper series j in year t that refer to projects published in journal or working paper series k between years $t-1$ to $t-6$; the same sample window to calculate a_{jt}.

Since Y_{it} is extremely right skewed, we take the logarithm of the output plus one to define our final time varying output measure:

$$y_{it} = \log(Y_{it} + 1).$$

Figure 1.6 plots the histogram average output level across 6 years for all sample advisees:

$$\bar{y}_i = \frac{1}{6} \sum_{t=0}^{6} y_{i,T_i+t}$$

where T_i denotes the year of graduation. This is the main output measure we use as the dependent variable for the production function and report in section 1.4. Note that, by definition, $\bar{y}_i$ covers all the output an advisee has produced 5 years before graduation and 5 years after graduation but with most information around the center which is one year after graduation. This allows us to capture the preliminary work done in graduate school but also work done as an independent scholar of an advisee, with an emphasis on the work that is likely most influenced by their advisor.

Pool of Choices – Advisors and Projects

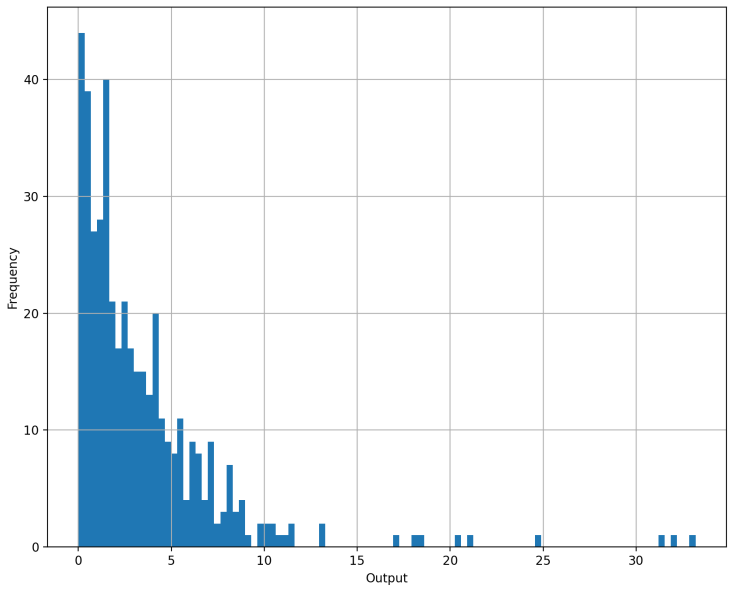
For each network growth model, we allow the advisee to make choices from a pool of choices instead of all possible choices, at least within the data. That is, for the genealogy network, the advisee chooses an advisor-author and for the coauthorship network, they choose a project from the corresponding pool of choices, instead of all advisors and all projects. This assumption is not only realistic to some extent, but is also keeps the model compact and reduce estimation noise.

We define the pool of choices for each case based on the graduation year and institution of each advisee as the following.

[1] *Pool of Advisors*. We assume that an advisee cannot choose an advisor outside of their institution thus we look at all the advisors for advisees from the same institution. Then for each advisee, we limit the pool of advisors to those who were advisors for the advisees who graduated within the past 5 year window, including their own.[20] The caveat is that we cannot

[20]Robustness test with 7 years, 10 years show no significant difference in outcome.

Figure 1.6: Histogram of the Average Output Level Across 6 Years after graduation – $\bar{y}_i$



Note: Out of 431 observations, even with the log transformation, the output values are exteremly skewed.

rule out the case of advisors who left or retired from the institution nor those who newly joined but haven't advised any advisee within the 5 year window.

[2] *Pool of Projects*. We define a cohort for each advisee, namely the advisees who graduated the year before, same year, and the year after from the same institution. Then the pool of projects are all the projects of those cohorts, except the ones that they are the sole author of, which publication or posting year is between 3 years before graduation to 1 year post graduation. The idea behind this is that if the cohort of an advisee participated in a project, it is likely that the advisee is capable of doing such as well. The caveat is that the strict compactness doesn't allow any potential outside projects, but given the fact that the time of when these projects are produced is during the advisees' graduate study, restricting the set is

Table 1.1: Summary Statistics for Each Pool to Selection Ratio

	Mean	Median	Variance	Skewness	Kurtosis
Pool of Advisors	0.1759	0.1429	0.1233	1.7458	4.9368
Pool of Projects	0.1463	0.0909	0.1525	2.4811	8.5110

Note: These summary statistics are for the sample of the following ratio for each advisee: the number of true choices over the number of potential choices. For example, each advisee would have a choice of roughly 6 potential advisors to choose from, as per the mean and median. Likewise, assuming the advisee only participates in one project, they have roughly 9 or 10 potential projects to choose from. Compared to the pool of advisors, the pool of projects are much skewed among advisees, mostly due to the difference in research activity among institutions.

not entirely unacceptable.

Table 1.1 shows the summary statistics of pool to selection ratio, which is the number of choices made of the total size of the pool for each advisee. Thus the for the pool of advisors, it's simply the inverse of the size of the pool, whereas for the pool of projects, it's the number of projects each advisee participated in during the defined window over the total possible projects they could've participated in.

Fixed Effect Homophily Variables

We collect the gender and region of origin information from *Gender-API.com* and use it to construct variables to account for the homophilic preferences for each network growth model. For each author we search for on *Gender-API.com*, the API returns a JSON file with the information of gender and country of origin. For gender, it returns a binary string value that indicates the gender (*male* or *female*) and the corresponding probability coined with the name that was used to search for. From this, we record the given gender which exceeds probability and remove those samples

Table 1.2: Number of Advisee, Advisor, and Other Authors Per Gender and Region of Origin

	Advisee	Advisor	Other Authors
No. Obs	431	457	115
Male	338	407	93
Female	93	50	21
Eastern Asia	71	13	16
Eastern Europe	68	50	10
Northern America	148	264	49
Northern Europe	64	128	19
South America	44	22	9
Southern Asia	39	21	15
Southern Europe	135	108	48
Western Asia	57	29	6
Western Europe	112	184	31
Other Regions	65	64	15

Note: Other authors are those that are neither an advisee nor an advisor, but authors for projects in the pool of projects. For the region of origin, each author has up to two categories so the numbers do not add up to the no. of total observations.

with *unknown* gender (probability 0.5)[21]

For the country of origin, the API returns a list of countries with positive probability greater than 0.01. A sample result would be in form of $\{$*USA* : 0.84, *Germany* : 0.54, *Denmark* : $0.08\}$, from which we take the two countries of the highest probability, i.e. *USA* and *Germany*, and use the statistical region – *North America* and *Western Europe* – as defined by Gender-API.com as the region of origin information for each author.[22] Note that results for some names returned only one country so those authors were labeled with one region instead of two.

Table 1.2 reports the number of each advisee, advisor, and other authors – those who are coauthors of projects in all pool of projects but not advisor-authors – along the corresponding gender and region of origin category. Note that one author can have up to two region of origin categories thus the total count of would not add up to the total number of observations.

Given the labels for each author, in the genealogy network growth model, we measure the degree of homophily by constructing pairwise dummy variables which measures the similarity between te advisee and advisor. To measure gender similarity, we define a dummy variable which takes value 1 if the advisee is the same gender with the advisor-author and 0 if not. Similarly, to measure the region of origin similarity, we define a dummy variable which takes value 1 if the set of region of origin of the advisee shares at least one region of origin (out of the two labels each author has) with the advisor author and 0 if not.

For the coauthorship network growth model, each project can have more than one author, so we don't use a binary variable. Instead, from the group of authors of each project, we collect the categorical information of the rest of the authors who are potential coauthors for each advisee. Then we calculate the ratio of those with the same homophilic characteristics with the advisee. For example, to constuct the variable that measures gender similarity, we calculate proportion of those who have the same gender from the original main authors (excluding her cohort who was

[21]Only 19 out of 12,635 that we searched for were *unknown*. The average correct probability was 0.962 for male and 0.934 for female.

[22]*Gender-API.com* provides three levels of granularity: country, statistical region, continental region. The statistical region is the second level which categorizes 234 countries in 22 groups. Detailed category information can be found at https://gender-api.com/en/api-docs/v2

the original participating advisee). Thus, for one of a female advisee's potential project which has one male original author and two female original authors, the advisee and project pairwise variable would take value of $2/3$. Similarly for the region of origin similarity variable, we calculate the proportion based on how many original authors share at least one region of origin with the advisee. Hence if one out of two original authors share at least one same region of origin, the advisee and project pairwise variable would take value $1/2$.

Given the two type of variables: i) advisee advisor-author pairwise variables and ii) advisee project pairwise variables, for each case, we construct the advisee fixed effect dummy variables that takes value for 1 for each category and 0 otherwise in each variable. That is, for the gender of advisee, we make a male and female dummy variable separately, eaching taking value of 1 for each corresponding gender and 0 otherwise. Using this, we can construct the interaction term by mulitplying these to similarity variables defined above. Then we can obtain the partial effects of gender similarity for the given gender of the advisee. Similarly for the region of origin, we can obtain the interaction term for each category that the advisee belongs to, though we only use the category with the higher probability of the two.

1.4 Results

In this section, we present the estimated results and follow up on a counterfactual study based on synthetic data generated from the two fitted network growth models.

1.4.1 Network Growth Models

Genealogy Network

Table 1.3 illustrates the estimated results for the genealogy network growth model for which we run three series of regressions through maximum likelihood estimation, where the likelihood is defined as in section 1.2.1. The reported standard errors in parentheses are based on numerical approximations of the hessian matrix. Regression (1) is the base line, which can be see as the raw preferential attachment model, whereas regression (2) includes the gender homophily variables and

regression (3) includes the region of origin homophily variables as well.

The significant estimates on the positive effect of past number of students on the network formation suggest that the genealogy network observes a preferential attachment behavior. Given such, adding the gender and region of origin homophily variables as in model (2) and (3) not changing the estimates much, suggest that even controlling for the homophilic factors, the tendency for new advisees to attach to advisors with more past students is prominent.

For gender homophily, we can see that the estimates are both positive for male and female advisees, suggesting that gender homophilic preference is observed, though it is less significant for the case of female advisees compared to the male advisees. This is mainly due to the lack of observations of female advisees (total 93) compared to that of male advisees (total 338). On the other hand, none of the region of origin homophily variables show statistical significance, which suggest no evidence of such homophily in the genealogy network.

In order to measure the goodness of fit of the estimation, we calculate the accuracy as the output is categorical. In a typical conditional logit model, if the choice set is homogeneous across all observations, the baseline accuracy is easily calculated – it is simply the probability of choosing one out of the total number of the choice set. However, in this setting, each advisee has a heterogeneous choice set, thus the baseline accuracy criterion is not equivalent for each advisee. Thus we calculate the accuracy of the case where we draw random advisors for each advisee from their corresponding pool as the baseline. The values in Table 1.3 are calculated based on 10,000 draws each.

In comparing the accuracy values for each regression model, we observe that all three models have at least a higher accuracy compared to the random baseline case, but not by a wide margin. The gap of the accuracy values across each model is narrower, which suggest that the homophilic preferences do not play a strong role in predicting the formation of a new advisor-advisee relationship as the information of the number of past students.

Coauthorship Network

Table 1.4 shows the estimated results for the coauthorship network growth model. Similar to the genealogy network growth model, we obtain the estimates via maximum likelihood estimation

Table 1.3: Estimated Results for Genealogy Network Growth Model

Variable	(1)	(2)	(3)
no. students	0.0301^{***}	0.0294^{***}	0.0288^{***}
	(0.0050)	(0.0051)	(0.0051)
male–male		0.4307^{*}	0.4335^{*}
		(0.2368)	(0.2376)
female–female		0.4138	0.4263
		(0.3117)	(0.3138)
Eastern Asia			0.2029
			(0.5908)
Eastern Europe			-0.214
			(0.4388)
Northern America			0.4008
			(0.2571)
Northern Europe			-0.2977
			(0.6345)
South America			0.1108
			(0.7256)
Southern Asia			0.2353
			(0.9867)
Southern Europe			-0.0036
			(0.3569)
Western Asia			0.2485
			(0.4495)
Western Europe			-0.0781
			(0.2823)
Other Region			0.9444
			(0.8698)
No. obs	431	431	431
Model Acc.	18.57%	18.70%	18.88%
Rnd. Acc.	17.58%	17.58%	17.58%

*p<0.1; **p<0.05; ***p<0.01

Note: Standard errors in parentheses are based on the approximated hessian matrix from the MLE estimation. Results suggest a clear pattern of preferential attachment from the significancy on the number of students, while weak evidence for gender homophily. Larger standard errors on the female-female coefficient is due to a smaller sample size compared to that of the male-male.

with the likelihood probability defined in section 1.2.2. Likewise, the reported standard errors in parentheses are obtained by numerical approximations of the hessian matrix. The number of observations for this model is the total number of projects done by all advisees.

We report 4 different regression models, where the baseline – regression (1) – is a model with a single variable; a dummy variable which indicates whether the project advisee participates is a single authored project or not. As we can see, the values are relatively similar across all 4 models which suggests that, unless given pairwise conditions are met, an average advisee would more likely to participate in a single authored paper compared to a coauthored paper.

The rest of the estimated models add the advisor based information, regression (2), or homophiliy preference variable, regression (3), or both – regression (4). The three new variables in model (2) are, for each advisee, i) a dummy variable, which takes value 1 if at least one of the coauthors is their advisor, ii) a dummy variable which takes value 1 if at least one of the coauthors is their advisor's past coauthor, and iii) a dummy variable, which takes value of 1 if the coauthor is another faculty member – this comes from the pool from section 1.3.2. We can see that the advisor related variables are statistically significant and also positive which suggests that advisees tend to work with their advisors or advisors' coauthors than other faculty members.

Regression (3) tests whether the homophilic preference have significancy for advisees making decisions which we see that, except for several region of origin variables, most are statistically insignificant. Especially, the male gender homophily coefficient is nearly zero and insignificant, suggesting that male advisees have no gender preference when selecting projects. On the other hand, the positive sign and significancy at a 20% level for female gender homophily coefficient suggests weak evidence for female advisees preferring to collaborate with other female coauthors than male coauthors.

Regression (4) includes all variables, where we see the significant estimates of advisor related variables from models (1) and (2) are consistently significant. On the other hand, the estimates on the gender homophily variables overall declined which suggests the likelihood of choosing the alternative choice regarding gender could be partially due to the advisor of the advisee being the same gender.

Table 1.4: Estimated Results for Coauthorship Network Growth Model

Variable	(1)	(2)	(3)	(4)
Not Single Authored	-1.5687^{***}	-1.8379^{***}	-1.600^{***}	-1.7787^{***}
	(0.0819)	(0.0915)	(0.1141)	(0.1185)
Advisor		1.3909^{***}		1.4608^{***}
		(0.1154)		(0.1196)
Advisor's Coauthor		0.6024^{***}		0.5830^{***}
		(0.1228)		(0.131)
Other Faculty		-0.9256^{***}		-0.8535^{***}
		(0.1496)		(0.1526)
Male			-0.0577	-0.1709
			(0.1139)	(0.1221)
Female			0.3632	0.1964
			(0.2362)	(0.2504)
Eastern Asia			0.2028	-0.354
			(0.6131)	(0.7109)
Eastern Europe			0.6226	0.5575
			(0.6589)	(0.7552)
Northern America			-0.3705	-0.5766^{*}
			(0.2987)	(0.3173)
Northern Europe			-0.5545	-0.4209
			(0.7779)	(0.8911)
South America			-0.3519	-0.2175
			(0.8508)	(0.8774)
Southern Asia			1.9006^{***}	1.7758^{***}
			(0.5327)	(0.6338)
Southern Europe			0.7215^{***}	0.9069^{***}
			(0.2980)	(0.3075)
Western Asia			1.1625^{***}	0.5263
			(0.5019)	(0.5363)
Western Europe			0.5549^{***}	0.0614
			(0.2769)	(0.3077)
Other Region			0.9284	0.7924
			(0.6133)	(0.6787)
No. obs	1114	1114	1114	1114
Model Acc.	10.55%	14.80%	11.39%	15.05%
Rnd. Acc.	10.78%	10.78%	10.78%	10.78%

*p<0.1; **p<0.05; ***p<0.01

Note: Standard errors in parentheses are based on the approximated hessian matrix from the MLE estimation. Results suggest advisees prefer to work on a single author project rather than coauthoring, though if they do coauthor, it is likely to be in close proximity with their advisor. Controlling this phenomena, we find no evidence for gender homophily though there are some region of origin homophily observed.

We report the accuracy of the models in the same manner as in section 1.4.1. We observe that models (2) and (4), namely, the ones with the advisor related variables, have a larger gap of increased accuracy over random selection compared to models without – (1) and (3). Especially, while the difference between model (1) and (3) is less than that of model (2) and (4), which suggest that homophilic preferences have a weak predicting power in the project selections process of advisees.

1.4.2 Production Function

In this section, we report the results of the production function estimation. We first estimate the mean regression of the log-linear model[23]. Then, we conduct a series of quantile regressions due to our interest in the overall distribution of output. The dependent variable for both regressions is the log of 6 year average of the output measure we defined in section 1.3.2 –$\bar{y}_i$ – including the year of graduation of each advisee.[24] [25]

In table 1.5, we report 5 models for the mean model, where each model includes institution and region of origin fixed effects; we omit to report due to most of the estimates being statistically insignificant.[26] Heteroscedasticity robust standard errors are reported in parentheses.

In the first two models, we investigate the effect of advisee and advisor gender on the average output. We find that, in our model, there is no statistical evidence of difference in output gender, nor the cases of advisees having same gender advisors.[27] This contradicts the works of numerous studies such as Ductor, Goyal and Prummer (2021) though this maybe due to our findings focusing on only the early stages of research, while the former finds significance evidence for established researchers throughout their career.[28] Also, contradicting to Gaule and Piacentini (2018), we show

[23]Estimated results for a linear model is in Table A.3. The results do not differ much.

[24]In detail, $\log\left(\frac{1}{6}\sum_{t=0}^{5} y_{i,T_i+t}\right)$.

[25]We conduct a study on each year after graduation up to 5 as well, which the results are in the Appendix.

[26]Full results are in the Appendix.

[27]Consistent with Hilmer and Hilmer (2007).

[28]We do find statistically significant difference in a linear model as shown in Table **??** in the Appendix, though it is less significant when controlling for the network statistics.

Table 1.5: Estimated Mean Regressions Results for the Log-Linear Production Function

	Dependent variable: $\log \bar{y}_i$				
	(1)	(2)	(3)	(4)	(5)
Advisee Male	0.221	0.277	0.165	0.227	0.195
	(0.159)	(0.534)	(0.508)	(0.511)	(0.155)
Advisor Male		0.197	0.126	0.162	
		(0.506)	(0.482)	(0.488)	
Both Male		-0.061	0.027	-0.036	
		(0.568)	(0.541)	(0.542)	
1st order Degree Centrality			2.101***	2.587***	2.611***
			(0.354)	(0.600)	(0.611)
2nd order Degree Centrality				-0.394	-0.413
				(0.467)	(0.474)
constant	1.084*	0.894	0.593	0.524	0.680
	(0.562)	(0.734)	(0.704)	(0.705)	(0.540)
Observations	431	431	431	431	431
R^2	0.294	0.295	0.332	0.333	0.333
Adjusted R^2	0.184	0.181	0.221	0.221	0.224

*p<0.1; **p<0.05; ***p<0.01

Note: Standard errors in parentheses are heteroscedasticity robust standard errors. Centrality measures are from the academic social network. We see clear positive correlation between the 1st order degree centrality network statistics across all models. Extended results are in Table A.2

that there is no evidence of having an advisor of same gender leading to higher research output, at least in the early stages of research for economics Ph.D. students.

In models (3) and (4), we include the network statistics from the academic social network, namely, the first and second order weighted degree centrality of the each advisee.[29] We find that the 1st order degree centrality is statistically significant as in model (3) and (4), but not the 2nd order degree centrality. Though insignificant, we observe a negative effect of the 2nd order weighted degree on average output.[30] These result suggests that starting with a larger volume of coauthored projects, and consequently coauthors, have a positive impact on early stage research, but also staying in a relatively smaller network, that is, having connections with less connected authors implies higher productivity.

Figure 1.7 illustrates the estimates of the coefficient on the 1st and 2nd order degree variable for a series of quantile regressions. Specifically, we use quantile parameter τ to be from 0.1 to 0.9, on the same variables as in model (5) in Table 1.5 as it has the highest goodness of fit based on the adjusted R^2 values. In both figures, the bar plots plot the estimate and 95% confidence intervals while the solid red line plots the estimates in model (5) and the dashed red lines plots its 95% confidence interval values.[31]

As we see in Figure 1.7a, given the control variables, the output difference in gender is mostly statistically insignificant, similar to that found in the mean regression models.[32] We observe similar results in Figure 1.7c for the 2nd order degree as well, where most have statistically insignificant – at 5% level – negative estimates except the coefficient for $\tau = 0.9$. On the contrary, the estimates for the 1st order degree centrality is statistically significant over all quantile levels. Moreover, we see that the estimated outputs for the lower and higher quantiles are more sensitive to the network statistic, compared to those in the middle range.

---

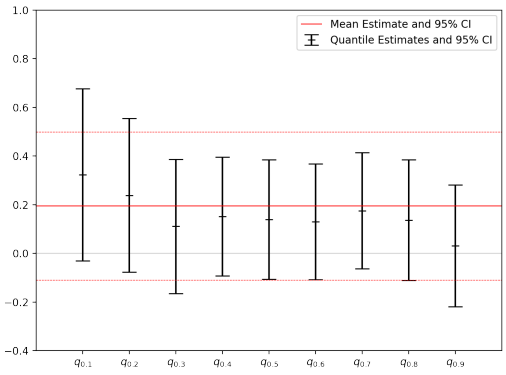
[29] All network statistic values were multiplied by 10000 for scaling purposes

[30] For higher order network statistics, Ductor, Goyal and Prummer (2021) find that the clustering coefficients are negatively correlated with research output.

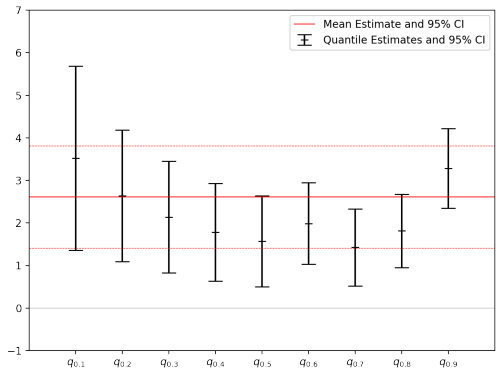
[31] Estimates for all other variables are in the Appendix.

[32] Only the estimate in $\tau = 0.1$ is statistically significant at the 10% level.

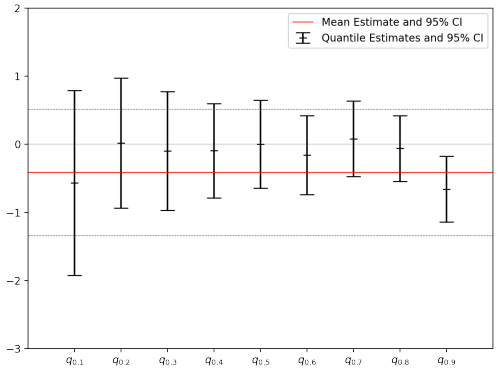
Figure 1.7: Estimated Quantile Regressions Results for Selected Variables over $\tau \in \{0.1, ..., 0.9\}$



a Advisee Male



b 1st order Degree Centrality



c 2nd order Degree Centrality

Note: Positive correlation between the 1st order degree centrality network statistics and output is observed across all quantiles. No significant difference by gender across all quantiles. Numerical Reports are in Table A.4.

Table 1.6: Network Generation Method for Each Case

	Case 1	Case 2	Case 3
Genealogy	Prediction	Random	Random
Coauthorship	Prediction	Prediction	Random

1.4.3 Counterfactual Study

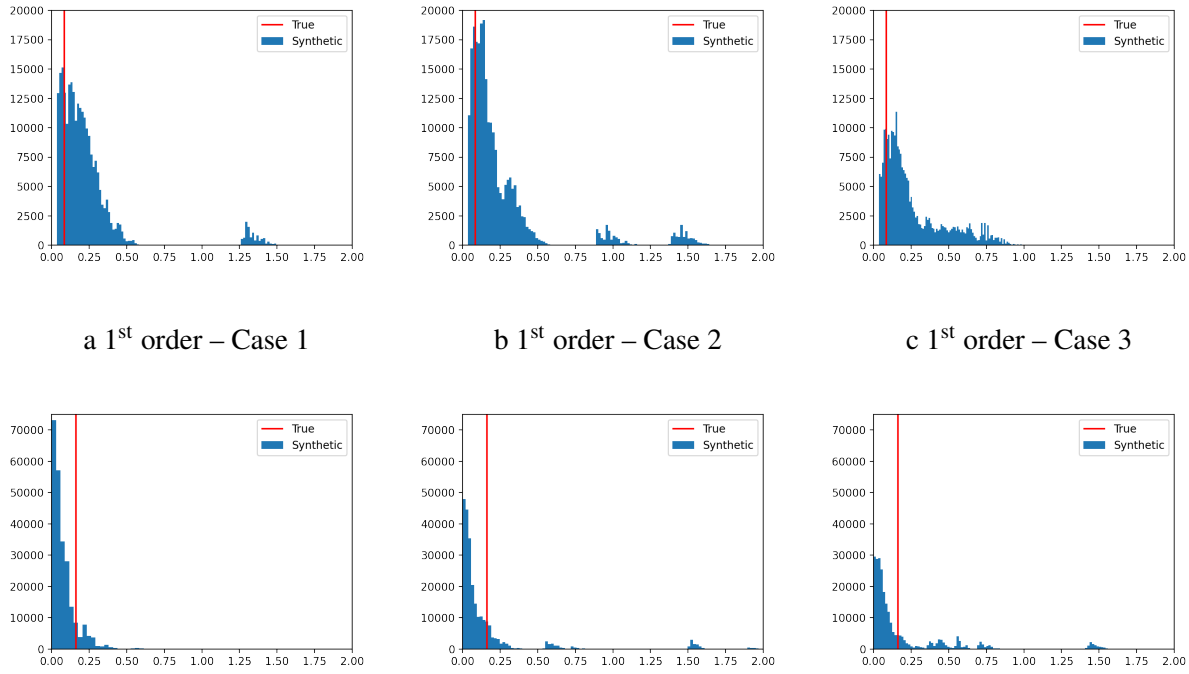
The counterfactual study aims to find the role of the two network growth models by translating the effect into the output measure. Thus we conduct the study based on synthetic network data generated by each network formation model, then run it through the production function.

For each network model, we choose the best performing model in terms of accuracy – model (3) for the genealogy network and model (4) for the coauthorship network – and generate synthetic networks by simulation, following the steps in section 1.2.4. Namely, we generate 500 genealogy networks and for each simulated genealogy network, we generate 500 synthetic coauthorship networks and calculate a total of 250,000 predicted output values using model (5) of the production function and the corresponding quantile regression models.

Given the data generation process above, we compare three cases as defined in Table 1.6. Case 1 is the case described in section 1.2.4, where advisors are sampled proportionately to the predicted probabilities and given such draws, the projects are sampled proportionately to the predicted conditional probabilities, both over their corresponding pool of choices. Case 2 is where the advisors are sampled randomly, that is proportionate to a uniform distribution over the pool of advisors, and given those samples, the projects are chosen by the conditional probabilities. Case 3 is where both advisors and projects are sampled randomly.

The difference between Case 1 and 2 provides insight on how advisor allocation on advisee effect their predicted productivity. By comparing the output from a random allocation of advisors to a advisee preference based one, we can account how much the advisor selection process takes role in predicting an advisee's research output. Case 3 goes a further step, where everything is randomly allocated thus providing a base line for our comparisons.

Figure 1.8: Sample of Generated Synthetic Network Statistics – Degree Centrality



a 1st order – Case 1 b 1st order – Case 2 c 1st order – Case 3

d 2nd order – Case 1 e 2nd order – Case 2 f 2nd order – Case 3

Note: Sample from one advisee. Top row: Simulated 1st order degree centrality values for each case. For roughly 78% of the cases, the median value is above the true centrality. Bottom row: Simulated 2nd order degree centrality values for each case. For roughly 40% of the cases, the median value is above the true centrality.

Table 1.7: Average of Individual Percentage Gains on Model Predicted Output and Among Cases for Each Case of Synthetic Data based Output

	% Gains on Model Predicted Output			% Gains Among Cases		
	Case 1	Case 2	Case 3	Case 1–2	Case 1–3	Case 2–3
Mean	0.5684	0.5661	0.5550	0.0326	0.0355	0.0303
$q_{0.1}$	0.8714	0.8715	0.8467	0.0575	0.0621	0.0549
$q_{0.2}$	0.6620	0.6487	0.6490	0.0540	0.0572	0.0463
$q_{0.3}$	0.4731	0.4656	0.4638	0.0314	0.0332	0.0261
$q_{0.4}$	0.3734	0.3675	0.3663	0.0223	0.0234	0.0178
$q_{0.5}$	0.3303	0.3235	0.3244	0.0206	0.0213	0.0154
$q_{0.6}$	0.4200	0.4144	0.4116	0.0250	0.0265	0.0208
$q_{0.7}$	0.3050	0.2975	0.2998	0.0202	0.0205	0.0142
$q_{0.8}$	0.3864	0.3796	0.3791	0.0242	0.0253	0.0192
$q_{0.9}$	0.7603	0.7628	0.7394	0.0443	0.0485	0.0435

Left Panel: Mean values of individual percentage gain from the true data predicted output for each model across each case. Roughly 55% gain of the mean model is due to the higher network statistics values than true for all cases. Case 1 is the highest for the Mean model and most quantiles.

Right Panel: Mean values of individual percentage gain from each predicted output of synthetic data cases across each model. Positive values across all rows and columns imply that predicted output based on predicted network formation improves upon random formation everywhere.

Figure 1.8 plots histograms of synthetic network statistics – 1st and 2nd order degree centrality – for each case, that of a sample advisee. The solid red line is the true statistic value for the corresponding advisee. For the synthetic 1st order degree centrality draws, we observe that, for roughly 97% of the 431 advisees, the true value is less than the mean of the generated draws, and for roughly 78%, the true value is less than the median, similarly for all three cases. For the synthetic 2nd order degree centrality draws, the proportion of those with respect to the advisees are 80% and 40%, respectfully.[33]

We first report the results on the individual level gain as in Table 1.7. The figures in the left

[33]Exact proportions are in Table A.1 of the Appendix.

panel are the average of individual percentage gain from the true data predicted output to synthetic data predicted output for each case. Note that the gains are around 55% in average, which is due to the high proportion of draws of the network statistics being greater than the true values.[34] Therefore, we compare the gains across each case, where we can see that the average gain values for Case 1 are the largest, except for the first and last quantile – even which the difference is negligible compared to Case 3.

For robustness, we also compare the average individual gains among cases, which is reported in the right panel of Table 1.7. We can see that the first two columns being positive on all models support the findings in the left panel, where Case 1 has the highest average individual gains. Moreoever, unlike how the Case 2 had higher gains to the model predicted output in the first and last quantile, the comparsion between Cases show that Case 1 has gains over both Cases. Thus, for each individual, in average, the predicted output with both model based synthetic data has a gain over the predicted output with synthetic data from either process being uniform random.

Next, we report the results of the predicted output values on the aggregate level in Table 1.8. Each value is the average of the predicted output for each model across the three cases. To compare Case 1 and 2, the results from the mean regression model show that the average predicted output of Case 1 is smaller, but not as much as a difference there is with respect to Case 3. For the results from quantile regressions, the average predicted output for Case 2 of the lower quantiles ($q_{0.1}, q_{0.2}, q_{0.3}$) and the highest quantile is larger but smaller for the middle quantiles. This suggests that it is difficult to conclude on whether the advisor allocation based on advisee preferences improves upon random allocation at the aggregate level.

The interesting result is that Case 3 dominates both Cases in all models in terms of average predicted output. This implies that, in the aggergate level, advisee-preference based allocation of advisor and projects are less efficient than that of the case of total random allocation. When comparing this result with the individual level gains, we can see that, in average, each advisee would be better off when they choose their prefered advisor and project, though the overall total

[34]The similar gain amount across all three cases suggests that it is likely due to the low accuracy of the fitted network growth models.

Table 1.8: Average Predicted Output ($\hat{\tilde{y}}_i$) for Each Model across Each Case

	Case 1	Case 2	Case 3
Mean	3.5510	3.5564	3.5999
$q_{0.1}$	1.3673	1.3702	1.4463
$q_{0.2}$	1.8886	1.9063	1.9255
$q_{0.3}$	2.3413	2.3499	2.3800
$q_{0.4}$	3.0029	2.9900	3.0180
$q_{0.5}$	3.4141	3.3911	3.4169
$q_{0.6}$	4.6166	4.5855	4.6175
$q_{0.7}$	5.4998	5.4735	5.5110
$q_{0.8}$	7.0593	7.0404	7.0647
$q_{0.9}$	12.7217	12.8037	12.8559

Average predicted output for Case 1 is greater than Case 2 for the mid level quantiles while less than the tail quantiles. Note that the predicted aggregate output for Case 3 dominate both cases everywhere. This implies the network formation through the predicted models result in less efficiency in terms of the aggregate productivity.

research output would be less than that of random allocation.

1.5 Remarks and Conclusion

As shown in section 1.4, we first find statistical significance of the genealogy network growth process to follow a preferential attachment-like formation model and find that there exists subtle gender homophily between the advisor and advisees. We also find that, advisees are more likely to join projects with their advisor or advisors' coauthor than cases where gender or region of origin are similar. Moreover, we discover that the network statistics from the *academic social network* are a viable proxy for early stage research output. Also, while there are output difference by gender of advisees, the gender of the advisors and it's match to advisees had no significant explanation power of early stage research output.

Through the counterfactual studies, we find some evidence that, compared to random allocation of advisors and advisees, advisee-preference based allocation allows advisees to gain more individ-

ual output levels. However, we also find that this would result in an overall lower aggregate output in average, but also across all predicted quantiles, suggesting that preference based allocation is less efficient than random allocation in the social planner's view. Given these findings, our novel attempt on measuring the allocation effect of advisors may shed some light to the areas of higher education and the informatics communitiy, but also policy makers within economics programs.

Chapter 2

THE IMPACT OF MISSING LINKS ON LINEAR REDUCED-FORM NETWORK-BASED PEER EFFECTS ESTIMATES

In this chapter, we analyze the implications of missing network links on reduced-form linear peer effects estimates. We show that linear-in-means estimates are generally attenuated, while linear-in-sums estimates may be attenuated, augmented, or consistent, depending upon the missingness mechanism and the relationship between the network and covariates. Our results rely on an assumption of "expectational equivalence", which implies that the covariance structure of the covariates between true network and the observed network does not differ. We derive expressions for the probability limits of estimators using observed network data and discuss special cases, including when estimators are consistent despite mismeasured networks in the linear-in-sums case. We demonstrate our results using two real datasets, including one with randomly-assigned covariates, and through simulations. The contrasting results for linear-in-means and linear-in-sums models suggest caution in extrapolating findings about measurement error across different peer effects estimators.

2.1 Intro

Research into network-based peer effects has exploded over the past decades (see Bramoullé, Djebbari and Fortin, 2020, for a recent summary).[1] The feasibility of studying the effects of peers in networks relies crucially on data that identifies who network peers are. However, such network data is often partially missing, leaving researchers to estimate these effects using the possibly-mismeasured *observed* network rather than *true* network that generated the data.

[1]Note that this is a more recent development than the classroom- or group-based peer effects literature that dates at least to Manski (1993), in which "peers" are defined as including all others within a given group or classroom (see Epple and Romano, 2011; Sacerdote, 2011, for reviews of this literature).

This paper analyzes the implications of this measurement error in commonly-used reduced-form linear peer effects estimators. To be concrete, a large number of studies estimate a form of the regression in Equation (2.1) (see Sacerdote, 2011; Bramoullé, Djebbari and Fortin, 2020).

$$y_i = \alpha_0 + x_i\alpha_1 + (\sum_{j \neq i} W_{ij}x_j)\alpha_2 + \epsilon_i \tag{2.1}$$

In Equation (2.1), y_i and x_i are individual i's outcome and covariate vector, while W_{ij} governs the influence that agent $j \neq i$ has on agent i's outcome. Common choices of weights W_{ij} correspond to well-known reduced-form linear models. With a link between agents i, j defined by the indicator L_{ij}, the linear-in-*means* model assumes $W_{ij} = \frac{L_{ij}}{\sum_k L_{ik}}$, while the linear-in-*sums* model assumes $W_{ij} = L_{ij}$.[2]

The parameters of interest are α_1 and α_2, the "direct" and "spillover" or "peer" effects, respectively.[3] An important special case arises in randomized trials where x_i is a randomly-assigned, binary treatment indicator. In such cases, α_2 identifies the marginal effect of either the average friends' treatment status (linear-in-means) or of additional treated friends (linear-in-sums). Even with standard exogeneity conditions, estimation may be problematic whenever W_{ij} is not perfectly observed. In such cases, the researcher instead estimates Equation (2.2), where W^*_{ij} is a mismeasured proxy for the true weights W_{ij}.

$$y_i = \alpha_0 + x_i\alpha_1 + (\sum_{j \neq i} W^*_{ij}x_j)\alpha_2 + \epsilon_i \tag{2.2}$$

We consider the common case in which only a subset of links are observed, either due to errors in reporting or explicit survey design (e.g., censoring). At first glance, the specification in Equation (2.2) resembles the canonical "classical measurement error" or "errors in variables" model, but we

[2]The linear-in-means and linear-in-sums models are also sometimes termed "local average" and "local aggregate" models, respectively, as discussed in Footnote 1 of Lewbel, Qu and Tang (2022) (see also Liu, 2013).

[3]Under appropriate conditions, α_1 and α_2 can be interpreted as reduced-form parameters of the standard linear-in-means/sums model that dates to Manski (1993), with the relationship between the "structural" parameters of known form in the group/classroom case (Manski, 1993; Carrell, Sacerdote and West, 2013) and given in Lemmas 4-5 herein for the network case.

show that the relationship to the well-known results depends crucially on the model estimated. Under stated conditions, we derive expressions for the probability limits of OLS estimators of α_1 and α_2 that use observed (rather than true) network data.

Our results rely upon a key assumption, which we define as "expectational equivalence." Essentially, we assume that the covariance between agents' own and their peers' covariates (in the true network) is, in expectation, the same for links that are observed and those that are not. That is, the set of *observed* links is representative (in terms of their covariates and networks) of the *true* links. A strength of our approach is that—other than moments existing—we make no assumptions on the network formation process. Further, links being observed (or missing) randomly is a sufficient—but by no means necessary—condition for our assumption to hold. Additionally, expectational equivalence is implied by random assignment, such as in field projects where the "covariate" is a binary treatment indicator.

While seemingly strong, our key assumption is weaker than others made in the literature, such as "order irrelevance" found in Griffith (2022). Similar to Lewbel, Qu and Tang (2022)—but in sharp contrast with Lewbel, Qu and Tang (2023*b*)—we make no assumption on the relationship between the network and covariates. Further, in contrast to Boucher and Houndetoungan (2019), we make no assumption about link independence across dyads.

We make two primary contributions. First, we show that the qualitative result from Griffith (2022)—that, in general, we should expect attenuation in estimated peer effects in the linear-in-means model—extends to the case with endogenous peer effects and with a weaker form of the key assumption. Further, the magnitude of inconsistency depends only on comparison of moments of the marginal distributions of true and observed degree. Indeed, under our assumptions, measurement error of this type in the linear-in-means model is a special case of "classical" measurement error.[4]

Second, we provide what is, to our knowledge, the first characterization of the effects of missing links in reduced-form estimation of the linear-in-sums model. These results are markedly different

[4]Strictly speaking, this is only true when x_i is scalar. See discussion in Subsection 2.3.3.

from the linear-in-means results: even under expectational equivalence, the inconsistency can be of arbitrary sign. That is, the probability limits of both direct and peer effect estimators can be too small, too large, or even consistent in some cases.

Given the theoretical ambiguity in the linear-in-sums case, we discuss additional structure that can help to alleviate this ambiguity. We focus on two types of restrictions: (i) on the relationship between covariates and the network; and (ii) on the "missingness" mechanism. To that end, we show that, when covariates are assigned randomly *and* links are missing randomly, the linear-in-sums estimator is consistent. A weaker restriction that we define as "weak homophily" gives further results.

These contrasting results may at first be surprising. However, linear-in-means and linear-in-sums models estimate conceptually different parameters. Linear-in-means models seek to identify the effect of peer *quality*, while linear-in-sums models identify the effect of peer *quantity*. What is being estimated matters, and results can be quite different even for conceptually very similar estimators. To further solidify this point, Lewbel, Qu and Tang (2022) show that, assuming random missingness, 2SLS estimates in the linear-in-sums model using the partially-observed network are biased upwards (in magnitude).[5]

While the linear-in-means model is more commonly used in practice, a number of recent works have estimated versions of the linear-in-sums model. For example, in a randomized experiment, Oster and Thornton (2012) study the effect of the number of treated friends on menstrual cup take-up. Bailey et al. (2022) study the effects of cell phone purchases on Facebook friends' cell phone purchasing decisions using a linear-in-sums specification.[6] Additionally, regression of any outcome on the number of links or *size* of one's social network is a special case of the linear-in-sums model.[7] Conti et al. (2013), Shi and Moody (2017), and Lleras-Muney et al. (2023)

[5]This is their benchmark case, but they also extend their results to allow missingness to depend on covariates. Our results here are more general in the sense that they allow for missingness to depend on degree.

[6]Further, in the context of geographically-defined networks, linear-in-sums specifications are found in, e.g., Miguel and Kremer (2004) (study peer effects in deworming) and Sampson and Perry (2019) (studying agricultural technology adoption). Brown and Laschever (2012) show the effect of an additional retirement at the same school on an individual's subsequent retirement decision, essentially estimating a linear-in-sums model.

[7]In this special case, $x_i = 1$ uniformly and thus Equation (2.1) is simply a regresssion of y_i on a constant and the

estimate effects of number of friends during adolescence on long-term outcomes. Other recent works have analyzed the effects of network size on subsequent incarceration among the unhoused (Corno, 2017), health outcomes among adolescents (Ho, 2016), fundraising participation (Scharf and Smith, 2016), and long-run outcomes of the next generation (Plug, van der Klaaw and Ziegler, 2018).[8] A related literature looks at the effect of network size on outcomes for immigrants (see, e.g. Beaman, 2011; Munshi, 2003). In sum, linear-in-sums models have the potential for wide application.

This paper is related to a number of other recent works. Griffith (2022) studies a special case of the linear-in-means estimator. Lewbel, Qu and Tang (2022) provide a treatment of the 2SLS strategy (see Bramoullé, Djebbari and Fortin, 2009) to recover the structural parameters of the linear-in-sums model, showing augmentation under conditions more restrictive than we assume here.[9] Relatedly, Boucher and Houndetoungan (2019) and Lewbel, Qu and Tang (2023*b*) both propose methods to recover structural parameters of linear-in-means models with limited or nonexistent data on the network.[10]

Other recent works have investigated the implications of mismeasured networks on subsequent regression outputs for centrality measures (Cai, 2022) and diffusion models (see Breza et al., 2020). Manresa (2016), de Paula, Rasul and Souza (2020), and Griffith and Peng (2023) propose methods of recovering the true network structure using repeated observations when the network may be completely unobserved. Chandrasekhar and Lewis (2011) and Hsieh et al. (2024) study the implications of missing node-level network data.[11] These results cover a wide variety of settings and

---

number of links. See Subsection 2.5.2.

[8]We note that these studies often employ IV or another strategy to address the endogeneity in network size, and thus are not directly estimating the specification that we describe above in Equations (2.1) and (2.2).

[9]In a related paper, Lewbel, Qu and Tang (2023*a*) analyze the implications of two-sided network mismeasurement on 2SLS estimates of linear-in-means models where the number of missing links is growing at a sufficiently slow rate compared to the sample size.

[10]In contrast to these works that seek to identify/estimate structural parameters of the linear-in-means model, our focus on reduced-form parameters simplifies the analysis somewhat and avoids well-known identification issues (see Manski, 1993; Blume et al., 2015; Bramoullé, Djebbari and Fortin, 2009; Lee, 2007).

[11]In a recent paper, Bramoullé and Maes (2024) study the implications of mismeasured *covariates*—in contrast to mismeasured networks—on linear-in-means estimates, as previously studied in Angrist (2014) and Sojourner

empirical strategies. Our results here suggest that the effects of missing network data can be quite subtle and depend on a number of factors, and that caution is warranted in extrapolating results to estimators of even quite similar parameters.

This paper proceeds as follows. In Section 2.2, we describe the model and what is observed. In Section 2.3, we define and discuss expectational equivalence, the main assumption behind our results. Section 2.4 provides our main results in a simplified setting in which covariates are independent across observations, and we present general results in Section 2.5. Section 2.6 demonstrates these results with simulated missing data from two real datasets. Section 2.7 concludes.[12]

2.2 Setup

In this section, we describe the assumed data generating process, definitions of weights, what is observed, technical conditions, and estimator definitions.

2.2.1 DGP

We assume that there exists a true network that is fully characterized by an $N \times N$ binary adjacency matrix $\mathbf{L}$.[13] From this adjacency matrix, a weighting matrix $\mathbf{W}$ is defined, where $\mathbf{W}$ determines how individuals influence each other. The outcome y_i for each individual i is determined according to the linear model in Equation (2.3).

$$y_i = \beta_0 + \beta_1 \sum_j W_{ij} y_j + x_i \beta_2 + \sum_j W_{ij} x_j \beta_3 + \epsilon_i \tag{2.3}$$

(2013) in the context of group-based interactions.

[12]Proofs of analytic results and a simulation study are included in the Supplemental Materials as Appendices C.1 and B.2, respectively.

[13]We also assume the absence of self-links and thus all diagonal elements of $\mathbf{L}$ are 0.

y_i, y_j are scalars and x_i, x_j are d-dimensional row vectors.[14] We rewrite Equation (2.3) in matrix form as Equation (2.4).

$$\mathbf{y} = \beta_0\iota + \beta_1\mathbf{W}\mathbf{y} + \mathbf{x}\beta_2 + \mathbf{W}\mathbf{x}\beta_3 + \epsilon \tag{2.4}$$

where ι is a vector of 1's.

Assumption 1. *We make the following technical assumptions on the data generating process:*

[1] $(\beta_1, \beta_2', \beta_3')' \in \mathbb{R}_{2d+1}$ *is a finite vector*

[2] For all i, $(x_i', \epsilon_i)' \in \mathbb{R}_{d+1}$, *where* $\mathbb{E}[x_i^4], \mathbb{E}[\epsilon_i^4] < \infty$.

[3] $\rho(\beta_1\mathbf{W}) < 1$.

We make technical assumptions that are maintained throughout. In short, Assumption 1 ensures that the system in Equation (2.4) is stable and that, with appropriate assumptions on independence, the conditions for the weak LLN are satisfied such that sample analogues of means and covariances converge in probability to true expectations. In part [3], the operator ρ defines the spectral radius, or largest eigenvalue (in absolute terms). This stability assumption ensures that we can employ the Neumann expansion $(\mathbf{I} - \beta_1\mathbf{W})^{-1} = \sum_{m=0}^{\infty} \beta_1^m \mathbf{W}^m$.[15]

2.2.2 Definitions of $\mathbf{W}$

We consider two different weighting schemes, and we further assume that the weighting scheme is known a priori.[16] Definitions 1 and 2 define the linear-in-means and linear-in-sums models,

[14] In the taxonomy of Manski (1993), β_1 and β_3 identify the "endogenous" and "exogenous" peer effects, respectively, while correlation among ϵ leads to "correlated" effects.

[15] Similar conditions are standard in the peer effects literature (see, e.g., Boucher and Houndetoungan, 2019; Griffith and Peng, 2023; Manski, 1993). In the linear-in-means model, the standard assumption that $|\beta_1| < 1$ implies part [3] of Assumption 1.

[16] But see de Paula, Rasul and Souza (2020), Manresa (2016), and Griffith and Peng (2023) for works that seek to estimate these weights.

respectively. The difference between the models is that the linear-in-means model has a row-normalized weighting matrix: this is performed by the $diag(\mathbf{L}\iota)$ term, where ι is an $N \times 1$ matrix of 1's. In contrast, in the linear-in-sums model, the row sum across $\mathbf{W}$ is simply each agent's degree. Both models feature prominently in the literature, as discussed in the introduction and in more detail in Bramoullé, Djebbari and Fortin (2020). To simplify notation, we define $\mathbf{G}$ as the (true) row-normalized adjacency matrix in Definition 1.

Definition 1. *In the linear-in-means model, outcomes are determined according to Equations (2.3)-(2.4) s.t.* $\mathbf{W} = (diag(\mathbf{L}\iota))^{-1}\mathbf{L} = \mathbf{G}$.

Definition 2. *In the linear-in-sums model, outcomes are determined according to Equations (2.3)-(2.4) s.t.* $\mathbf{W} = \mathbf{L}$.

2.2.3 What Is Observed

Assumption 2 details what is observed. In general, the true-data adjacency matrix $\mathbf{L}$ is not observed. Rather, we observe $\mathbf{M}$, another binary matrix that contains a subset of links. From this, we can construct $\mathbf{W}^*$, the observed-data analogue of $\mathbf{W}$.

Assumption 2. *The following are observed:*

[1] $\mathbf{y}$*, an* $N \times 1$ *vector of demeaned outcomes*

[2] $\mathbf{x}$*, an* $N \times d$ *vector of covariates, where* d *is the dimension*

[3] $\mathbf{M}$*, an* $N \times N$ *binary adjacency matrix, where* $\mathbf{M}_{(i,j)} \leq \mathbf{L}_{(i,j)}$ *for all* i, j.

We also state an exogeneity assumption that is maintained throughout, given in Assumption 3. We note that this exogeneity assumption is stronger and weaker than others found in the literature. Bramoullé, Djebbari and Fortin (2020) point out that, in general, a weaker condition is sufficient for *identification* of the structural parameters in the linear-in-means (but not linear-in-sums) case. In contrast, Lewbel, Qu and Tang (2022) assume that, conditional on a link existing ($\mathbf{L}_{(i,j)} = 1$),

$\mathbf{M}_{(i,j)}$ is only conditional on covariates $\mathbf{x}$.[17] That is, they make a stronger assumption on the relationship between $\mathbf{L}$ and $\mathbf{M}$ than we make here, one that rules out some joint distributions, such as degree censoring that we discuss below.

Assumption 3. $\mathbb{E}[\epsilon_i|\mathbf{x}, \mathbf{L}, \mathbf{M}] = 0$

Finally, we state rank conditions that ensure invertibility of matrices and that, in turn, regression estimators are well-defined. These conditions are given in Assumption 4.

Assumption 4. *The matrices* $\mathbb{E}\left[\frac{1}{N}[\mathbf{x}, \mathbf{W}\mathbf{x}]'[\mathbf{x}, \mathbf{W}\mathbf{x}]\right]$ *and* $\mathbb{E}\left[\frac{1}{N}[\mathbf{x}, \mathbf{W}^*\mathbf{x}]'[\mathbf{x}, \mathbf{W}^*\mathbf{x}]\right]$ *have full rank* $(2d)$*.*

2.2.4 Estimator and Parameter Definitions

We study an estimator of reduced-form parameters of the model in Equations (2.3)-(2.4). First, we define $\hat{\alpha}^{true}$, the true-data reduced-form estimator, in Definition 3. This estimator is simply an OLS estimation of the outcome $\mathbf{y}_{(i)}$ on the vector of agent-i's own characteristics $\mathbf{x}_{(i)}$ and a weighted sum of all others' covariates, $(\mathbf{W}\mathbf{x})_{(i)}$.[18] Note that we do not estimate a separate constant in the regression, which simplifies the derivations and is without loss of generality.[19]

Definition 3. *The (true data)* <u>*RF Estimator*</u> $\hat{\alpha}^{true} = \left([\mathbf{x}, \mathbf{W}\mathbf{x}]'[\mathbf{x}, \mathbf{W}\mathbf{x}]\right)^{-1}[\mathbf{x}, \mathbf{W}\mathbf{x}]'\mathbf{y}$ *, where* $\mathbf{W}$ *is constructed from* $\mathbf{L}$*.*

In general, the estimator $\hat{\alpha}^{true}$ is infeasible since $\mathbf{M}$ (and not $\mathbf{L}$) is observed. Therefore, for a given weighting scheme, we construct the *observed data RF estimator* $\hat{\alpha}^{obs}$ as defined in Definition

[17] Our Assumption 3 is the same as their condition (A3). We need not impose their conditions (A1) and (A2), but note that we do require an additional condition on the relationship between covariates in the missing and non-missing links, given in Assumption 5.

[18] We adopt the convention that, for a given matrix (or vector) $\mathbf{V}$, $\mathbf{V}_{(i)}$ refers to row (or element) i. Therefore, $\mathbf{y}_{(i)}$ gives element i of vector $\mathbf{y}$, while $(\mathbf{W}\mathbf{x})_{(j)}$ gives row j of matrix $(\mathbf{W}\mathbf{x})$.

[19] That is, one can always simply demean prior to estimation which will not change the estimates, as implied by the Frisch-Waugh-Lovell Theorem. Alternatively, in the linear-in-sums case, a vector of 1's may be included as part of $\mathbf{x}$, in which case the corresponding coefficient of α_1 recovers the constant term while the corresponding coefficient of α_2 identifies the effect of number of friends. Versions of this have been estimated in practice, as discussed in the Introduction and Subsection 2.5.2 below.

4. The primary goal of this paper is to characterize the relationship between the probability limit of the feasible $\hat{\alpha}^{obs}$ and the infeasible $\hat{\alpha}^{true}$.

Definition 4. *The Observed-data RF Estimator* $\hat{\alpha}^{obs} = ([\mathbf{x}, \mathbf{W}^*\mathbf{x}]'[\mathbf{x}, \mathbf{W}^*\mathbf{x}])^{-1}[\mathbf{x}, \mathbf{W}^*\mathbf{x}]'\mathbf{y}$*, where* $\mathbf{W}^*$ *is constructed from* $\mathbf{M}$ *rather than* $\mathbf{L}$*.*

As in Griffith (2022), we adopt the definition of the reduced-form parameter α in Definition 5. In short, the reduced-form parameter is simply the probability limit of the true-data estimator given the assumptions above in the DGP. This can be thought of as a generalization of the reduced-form parameter definition from the classroom setting (where $\mathbf{G} = \mathbf{G}^2$) (see Manski, 1993; Carrell, Sacerdote and West, 2013).

Definition 5. *The reduced-form direct and peer effects parameter* $\alpha = \operatorname{plim} \hat{\alpha}^{true}$.

2.2.5 Missingness Mechanisms

To demonstrate intuition and our main results, we define two mechanisms through which a subset of the true network links may be observed/missing. We refer to these as "missingness mechanisms." The first, *random missingness*, randomly selects a predefined portion of the total number of edges of the true network. This method simplifies the analysis, as the relationship between the true and observed networks can be represented by a single parameter: e.g. $p = 0.5$, if 50% of the edges are unobserved randomly.

The second method, *degree censoring*, looks at only the first few connections reported by each person by a predefined threshold, a common survey design discussed in Griffith (2022). With this mechanism, the agents' degrees are "censored," which leads to complicated dependencies between true and observed degrees that does not have a simple functoinal form. Details on how the methods work in application is discussed in Section 2.6.

2.2.6 Asymptotic Framework

The results presented in the following sections make extensive use of the Weak Law of Large Numbers, premised on independence across groups $s = 1, ..., S$, where $S \to \infty$. Individuals are

in groups indexed by s, where each group consists of N_s agents such that $\sum_s N_s = N$ and who are connected through a network structure represented by adjacency matrix $\mathbf{L}_s$. That is, there is a sequence of true networks $\mathbf{L}_s$, with a corresponding sequence of observed networks, as follows:

True Networks	$\mathbf{L}_1$	$\mathbf{L}_2$	...	$\mathbf{L}_S$
Observed Networks	$\mathbf{M}_1$	$\mathbf{M}_2$	...	$\mathbf{M}_S$

Independence across networks leads to a block diagonal structure for $\mathbf{L}$ and $\mathbf{M}$. Together with $S \to \infty$, and the assumption of bounded fourth moments (as stated in Assumption 1), empirical variances/covariances converge in probability to the corresponding population moments .That is, for example,

$$\text{plim}\, \frac{1}{N} \sum_{s=1}^{S} \sum_{i=1}^{N_s} x_{is} x'_{is} = \lim_{S\to\infty} \frac{1}{N} \sum_{s=1}^{S} \sum_{i=1}^{N_s} x_{is} x'_{is} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i x'_i = \mathbb{E}[x_i x'_i] \tag{2.5}$$

In matrix notation, $\text{plim}(\frac{1}{N}\mathbf{x}'\mathbf{x}) = \mathbb{E}[x_i x'_i]$. To simplify notation, we drop the s subscript in the derivations and proofs.

2.3 Expectational Equivalence

2.3.1 Assumption Statement

In addition to Assumptions 1 - 4, our results require an assumption on the relationship between the links that are observed an those that are not. This is given in Assumption 5. Essentially, we assume that the variance structure in covariates of *observed* linked agents is the same in expectation as the variance structure in covariates of *all* linked agents (whether observed or not).[20]

Assumption 5. *(Expectational Equivalence) The following hold for all d^{true}, d^{obs}, $m = 0, ..., \infty$,*

[1] $\mathbb{E}[\mathbf{x}_{(i)}(\mathbf{W}^{\mathbf{m}}\mathbf{x})'_{(j)} | M_{ij}, L_{ij} = 1, d_i^{true} = d^{true}] = \mathbb{E}[\mathbf{x}_{(i)}(\mathbf{W}^{\mathbf{m}}\mathbf{x})'_{(j)} | L_{ij} = 1, d_i^{true} = d^{true}]$

[20] We assume (in Assumption 2) that the covariate vector $\mathbf{x}_{(j)}$ is observed for each agent j. Conditional on i having a link to j, it is possible that the *link* is not observed, and thus $\mathbf{x}_{(j)}$ would be included in $(\mathbf{W}\mathbf{x})_{(i)}$ but not in $(\mathbf{W}^*\mathbf{x})_{(i)}$.

[2] $\mathbb{E}[\mathbf{x}_{(j)}(\mathbf{W^m x})'_{(j)}|M_{ij}, L_{ij} = 1, d_i^{obs} = d^{obs}, d_i^{true} = d^{true}] = \mathbb{E}[\mathbf{x}_{(j)}(\mathbf{W^m x})'_{(j)}|L_{ij} = 1, d_i^{true} = d^{true}]$

[3] $\mathbb{E}[\mathbf{x}_{(j)}(\mathbf{W^m x})'_{(k)}|M_{ij}, M_{ik}, L_{ij} = 1, L_{ik} = 1, d_i^{obs} = d^{obs}, d_i^{true} = d^{true}] = \mathbb{E}[\mathbf{x}_{(j)}(\mathbf{W^m x})'_{(k)}|L_{ij} = 1, L_{ik} = 1, d_i^{true} = d^{true}]$

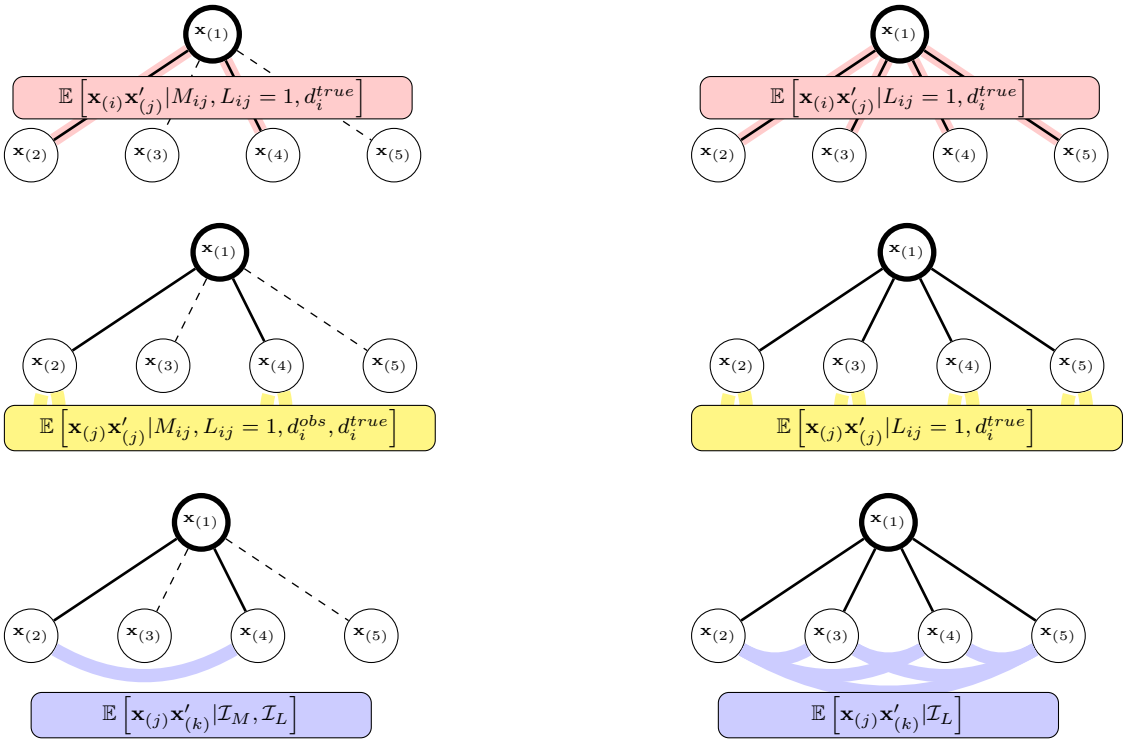
where i, j, k are distinct.

For each of parts [1]-[3] of Assumption 5, the covariance on the right-hand side of the equal sign is a feature of the true data-generating process. For example, in part [1] when $m = 0$, $\mathbb{E}[\mathbf{x}_{(i)}\mathbf{x}'_{(j)}|L_{ij} = 1, d_i^{true} = d^{true}]$ is simply the covariance between agents' covariate and linked agents' covariate, conditional on degree. These objects necessarily depend on the joint distribution of $\mathbf{x}_{(i)}$, $\mathbf{x}_{(j)}$ and the network (which determines L_{ij} and d_i^{true}). We only assume that these conditional covariances *exist*, and we make no further assumptions on the joint distribution of covariates and the true network.

Further to this point, we stress that Assumption 5 makes no assumptions on the relationship between *degree* and the covariance structure.[21] Rather, the covariances in the true network may vary arbitrarily by degree. For example, if agents with more friends also have more friends unlike them, $\mathbb{E}[\mathbf{x}_{(i)}\mathbf{x}'_{(j)}|M_{ij}, L_{ij} = 1, d_i^{true} = d^{true}]$ (or equivalently $\mathbb{E}[\mathbf{x}_{(i)}(\mathbf{x})'_{(j)}|L_{ij} = 1, d_i^{true} = d^{true}]$) would be decreasing in true degree. Assumption 5 allows for such dependence.

Rather than assumptions about the process that determines the network and covariates, Assumption 5 imposes assumptions on the relationship between the true and observed network, *conditional* on $\mathbf{x}$ and $\mathbf{L}$. That is, we only assume that, conditional on a link existing ($L_{ij} = 1$), conditional covariances are the same whether the link between i, j is observed ($M_{ij} = 1$) or not observed ($M_{ij} = 0$). This in turn implies that the covariances for observed links ($M_{ij} = 1$) are the same as those of all links ($L_{ij} = 1$).

[21] We note that earlier versions of this paper stated a stronger version of Assumption 5 that also conditioned on degree. This assumption did in fact impose restrictions on the relationship between the covariates and the network formation process, which Assumption 5 does not.



a Observed (Censored) Network (**M**)

b True Network (**L**)

Figure 2.1: Assumption 5 Implications. Note that $\mathcal{I}_M = \{M_{ij}, M_{ik}, d_i^{obs}\}$ and $\mathcal{I}_L = \{L_{ij} = 1, L_{ik} = 1, d_i^{true}\}$

We illustrate Assumption 5 by example, with Figure 2.1. In the true network, Agent 1 is linked to Agents 2-5. However, the researcher only observes the links to Agents 2 and 4 (denoted with solid lines), while links to 3 and 5 are not observed (denoted with dashed lines). We illustrate the three parts of Assumption 5 for $m = 0$, noting that the intuition is similar for $m > 0$. We stress that Assumption 5 only needs to hold *in the population* and that, due to sampling variability, is less likely to hold for individuals.

Assumption 5, part [1] implies that the expected covariance of the covariates between agent 1 and their observed connected agents – 2 and 4 – is the same as that of their true connected agents – 2, 3, 4, and 5. Similarly, part [2] implies that the expected variance of covariates of the neighbors of agent 1 is the same in both cases. Part [3] implies the expected covariance of the covariates among

the neighbors is the same in the observed and true network. These assumptions can be seen as implying that the variance structure—or expected value of the (cross) products—of the covariates among agents and their m-th order neighbors is equivalent in the true and observed networks.

By making assumptions only on the relationship between distriutions in the true and observed networks, we can provide some guidance to practitioners. While Assumption 5 appears complicated, it will hold whenever the (covariates of) observed links are representative of the true links that exist in the population. This in turn implies that *random missingness*—whereby observed links are an i.i.d. subset of the true links—is sufficient for the assumption to hold. Random missingness is not necessary, however, and Assumption 5 will hold in more general cases; for example, random assignment of the covariate also implies it.

2.3.2 Implications of Expectational Equivalence

Assumption 5 immediately leads to a number of intermediate results that highlight the difference between the linear-in-means and linear-in-sums cases. These results are then used in proving our results regarding the (in)consistency of $\hat{\alpha}^{obs}$.

Before proceeding, to simplify exposition, we define the following matrices, noting that each is a $d \times d$ finite matrix.

[1] $\mathbf{A}_m(d^{true}) = \mathbb{E}[\mathbf{x}_{(i)}(\mathbf{W^m x})'_{(j)} | L_{ij} = 1, d_i^{true} = d^{true}]$

[2] $\mathbf{B}_m(d^{true}) = \mathbb{E}[\mathbf{x}_{(j)}(\mathbf{W^m x})'_{(j)} | L_{ij} = 1, d_i^{true} = d^{true}]$

[3] $\mathbf{C}_m(d^{true}) = \mathbb{E}[\mathbf{x}_{(j)}(\mathbf{W^m x})'_{(k)} | L_{ij} = 1, L_{ik} = 1, d_i^{true} = d^{true}]$

These are defined for all $m = 0, ..., \infty$. We note that each of $\mathbf{A}_m(d^{true})$, $\mathbf{B}_m(d^{true})$, and $\mathbf{C}_m(d^{true})$ is a conditional covariance in the true network, as shown on the right-hand side of corresponding parts of Asssumption 5. Next, define $\mathbf{H} = (diag(\mathbf{M}\iota))^{-1}\mathbf{M}$, the observed-data analogue of $\mathbf{G}$.

First, Lemma 1 gives expressions for the covariance between agents' $\mathbf{x}$ and $\mathbf{Wx}$, their network neighbors' $\mathbf{x}$ (and higher-order terms). Note that expectations are taken over the distribution of

degree.[22] For example, when $m = 1$, Lemma 1, parts [1]-[2] imply that, in expectation, the covariance between Agent 1's covariates and the average of his/her observed neighbors' covariates (2 and 4) is the same as the covariance including all their neighbors (2-5, observed and unobserved).

Lemma 1. *Given Assumptions 1 and 5, for any* $m \geq 1$

If $\mathbf{W} = \mathbf{G}$,

[1] $\mathbb{E}[\frac{1}{N}\mathbf{x}'(\mathbf{G}^m\mathbf{x})] = \mathbb{E}[\mathbf{A}_{m-1}(d^{true})]$

[2] $\mathbb{E}[\frac{1}{N}\mathbf{x}'(\mathbf{H}\mathbf{x})] = \mathbb{E}[\mathbf{A}_0(d^{true})]$

If $\mathbf{W} = \mathbf{L}$,

[3] $\mathbb{E}[\frac{1}{N}\mathbf{x}'(\mathbf{L}^m\mathbf{x})] = \mathbb{E}[d^{true}\mathbf{A}_{m-1}(d^{true})]$

[4] $\mathbb{E}[\frac{1}{N}\mathbf{x}'(\mathbf{M}\mathbf{x})] = \mathbb{E}[d^{obs}\mathbf{A}_0(d^{true})]$

In contrast, Lemma 1, parts [3]-[4] show that the covariance between agents' $\mathbf{x}$ and the *sum* of his/her neighbors' covariates are in general not the same. Rather, the true covariance is larger (in magnitude) since it includes the sum across d^{true} agents' covariates, while the observed-data covariance calculates the sum across only $d^{obs} \leq d^{true}$ other agents' covariates.

Lemma 2. *Given Assumptions 1 and 5, for any* $m \geq 1$

If $\mathbf{W} = \mathbf{G}$,

[1] $\mathbb{E}[\frac{1}{N}(\mathbf{G}\mathbf{x})'\mathbf{G}^m\mathbf{x}] = \mathbb{E}[\mathbf{C}_{m-1}(d^{true})] + \mathbb{E}[\frac{1}{d^{true}}(\mathbf{B}_{m-1}(d^{true}) - \mathbf{C}_{m-1}(d^{true}))]$

[2] $\mathbb{E}[\frac{1}{N}(\mathbf{H}\mathbf{x})'\mathbf{G}^m\mathbf{x}] = \mathbb{E}[\mathbf{C}_{m-1}(d^{true})] + \mathbb{E}[\frac{1}{d^{true}}(\mathbf{B}_{m-1}(d^{true}) - \mathbf{C}_{m-1}(d^{true}))]$

[3] $\mathbb{E}[\frac{1}{N}(\mathbf{H}\mathbf{x})'\mathbf{H}\mathbf{x}] = \mathbb{E}[\mathbf{C}_0(d^{true})] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0(d^{true}) - \mathbf{C}_0(d^{true}))]$

If $\mathbf{W} = \mathbf{L}$,

[22] That is, for example, $\mathbb{E}[\mathbf{A}_0(d^{true})] = \sum_z Pr(d^{true} = z)\mathbf{A}_0(z)$.

[4] $\mathbb{E}[\frac{1}{N}(\mathbf{Lx})'\mathbf{L}^m\mathbf{x}] = \mathbb{E}[d^{true}(d^{true}-1)\mathbf{C}_{m-1}(d^{true})] + \mathbb{E}[d^{true}\mathbf{B}_{m-1}(d^{true})] = \mathbf{D}_m^{true}$

[5] $\mathbb{E}[\frac{1}{N}(\mathbf{Mx})'\mathbf{L}^m\mathbf{x}] = \mathbb{E}[d^{obs}(d^{true}-1)\mathbf{C}_{m-1}(d^{true})] + \mathbb{E}[d^{obs}\mathbf{B}_{m-1}(d^{true})] = \mathbf{D}_m^{int}$

[6] $\mathbb{E}[\frac{1}{N}(\mathbf{Mx})'\mathbf{Mx}] = \mathbb{E}[d^{obs}(d^{obs}-1)\mathbf{C}_0(d^{true})] + \mathbb{E}[d^{obs}\mathbf{B}_0(d^{true})] = \mathbf{D}_m^{obs}$

Next, Lemma 2 gives covariances between and among true- and observed-data means and sums. Parts [1]-[3] give results for the linear-in-means model, while [4]-[6] give results for the sums model. Importantly, note that [1] and [2] are the same: for example, when $m = 1$, $\mathbb{E}[\frac{1}{N}(\mathbf{Gx})'\mathbf{Gx}] = \mathbb{E}[\frac{1}{N}(\mathbf{Hx})'\mathbf{Gx}]$: the *variance* of the average of true peers' covariates is the same as the *covariance* between the average of true peers' covariates and the average of observed peers' covariates. Part [3], in contrast, shows that the variance of the average of observed-data peers' covariates is in general different. In fact, when $d = 1$, it is larger, since both $\frac{1}{d^{obs}} \geq \frac{1}{d^{true}}$ and $\mathbf{B}_0(d^{true}) \geq \mathbf{C}_0(d^{true})$ for all d^{true}.[23]

Parts [4]-[6] give analogous results for the linear-in-sums case. Note that, in general, for any m, [4] is larger in magnitude than [5]: that is, the variance of the sum of true-data peers' covariates is greater than the covariance between the sum of true-data peers' covariates and that of the observed-data peers. Finally, note that when $m = 1$, both of these are larger in magnitude than the variance of the observed-data sum, as given in part [6].

Lemma 3. *Given Assumptions 1, 3, and 5,*

If $\mathbf{W} = \mathbf{G}$*,*

[1] $\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] = \mathbf{E}_{\mathbf{xx}}\beta_2 + (\sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[\mathbf{A}_m(d^{true})])\,(\beta_1\beta_2+\beta_3)$

[2] $\mathbb{E}[\frac{1}{N}(\mathbf{Gx})'\mathbf{y}] = \mathbb{E}[\mathbf{A}_0'(d^{true})]\beta_2 + \sum_{m=0}^{\infty}\beta_1^m\Big(\mathbb{E}[\mathbf{C}_m(d^{true})] + \mathbb{E}[\frac{1}{d^{true}}(\mathbf{B}_m(d^{true}) - \mathbf{C}_m(d^{true}))]\Big)(\beta_1\beta_2+\beta_3)$

[3] $\mathbb{E}[\frac{1}{N}(\mathbf{Hx})'\mathbf{y}] = \mathbb{E}[\mathbf{A}_0'(d^{true})]\beta_2 + \sum_{m=0}^{\infty}\beta_1^m\Big(\mathbb{E}[\mathbf{C}_m(d^{true})] + \mathbb{E}[\frac{1}{d^{true}}(\mathbf{B}_m(d^{true}) - \mathbf{C}_m(d^{true}))]\Big)(\beta_1\beta_2+\beta_3)$

[23] The latter is due to the Cauchy-Schwartz Inequality.

If $\mathbf{W} = \mathbf{L}$*,*

[4] $\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] = \mathbf{E}_{\mathbf{xx}}\beta_2 + (\sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m(d^{true})])\,(\beta_1\beta_2+\beta_3)$

[5] $\mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{y}] = \mathbb{E}[d^{true}\mathbf{A}_0(d^{true})]'\beta_2 + \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m^{true}$

[6] $\mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{y}] = \mathbb{E}[d^{obs}\mathbf{A}_0(d^{true})]'\beta_2 + \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m^{int}$

where $\mathbf{E}_{\mathbf{xx}} = \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}]$.

Finally, Lemma 3 gives further results that build upon Lemmas 1 and 2. These results take advantage of the Neumann expansion to express $(\mathbf{I}-\beta_1\mathbf{W})^{-1}$ as a convergent series, as well as the exogeneity conditions stated in Assumption 3. Note that [2] and [3] are equivalent: the covariance between the average of peers' covariates and outcomes are the *same* in the true and observed networks. Contrast this with a comparison of [5] and [6], where the covariance between the *sum* of links' covariates differs between the true and observed network. These different intermediate results imply different patterns of inconsistency, as discussed in the following sections.

2.3.3 Relation to "Classical" Measurement Error Model

Lemma 2 allows us to make statements about the relationship between the true regressors ($\mathbf{W}\mathbf{x}$) and those measured with error ($\mathbf{W}^*\mathbf{x}$). To see this, write the observed-data peer covariate $\mathbf{W}^*\mathbf{x}$ as the sum of the truth $\mathbf{W}\mathbf{x}$ and error, as shown in Equation (2.6).

$$\mathbf{W}^*\mathbf{x} = \underbrace{\mathbf{W}\mathbf{x}}_{\text{true}} + \underbrace{(\mathbf{W}^*-\mathbf{W})\mathbf{x}}_{\text{error}} \tag{2.6}$$

In canonical "errors-in-variables" or "classical measurement error" models, the implications of measurement error depends on the covariance between the true measure and the error.[24] That is, it depends on $\mathbb{E}[\frac{1}{N}(\mathbf{W}\mathbf{x})'(\mathbf{W}^*-\mathbf{W})\mathbf{x}]$.

[24] Wooldridge (2010), Section 4.4.2, calls the assumption of no covariance the "classical errors-in-variables (CEV)" assumption. Cameron and Trivedi (2005) define it similarly and give a more general result that shows how measurement error of a single regressor may affect the consistency of parameters involving other covariates, as here when $d = 1$. Note that the canonical model generally makes statements about inconsistency when a *single* covariate

Assumption 5 allows us to state relatively simple, closed-form expressions for these covariances. For the linear-in-means model, we have $\mathbf{W}^* = \mathbf{H}$ and $\mathbf{W} = \mathbf{G}$. Therefore,

$$\mathbb{E}[\frac{1}{N}(\mathbf{W}\mathbf{x})'(\mathbf{W}^* - \mathbf{W})\mathbf{x}] = \mathbb{E}[\frac{1}{N}(\mathbf{G}\mathbf{x})'\mathbf{H}\mathbf{x}] - \mathbb{E}[\frac{1}{N}(\mathbf{G}\mathbf{x})'\mathbf{G}\mathbf{x}] = 0 \tag{2.7}$$

where Lemma 2 parts [2] and [3] imply that the expression in Equation (2.7) is 0! That is, in the linear-in-means model, the error does not covary with the true value. In this sense, the linear-in-means model conforms to the assumptions of the classical model. When $d = 1$, this implies attenuation of estimated coefficients of the peer effect parameter.

In contrast, the linear-in-sums model is quite different. Replace $\mathbf{W}^* = \mathbf{M}$ and $\mathbf{W} = \mathbf{L}$. So,

$$\begin{aligned}\mathbb{E}[\frac{1}{N}(\mathbf{W}\mathbf{x})'(\mathbf{W}^* - \mathbf{W})\mathbf{x}] &= \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{M}\mathbf{x}] - \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{L}\mathbf{x}] \\ &= -\mathbb{E}[(d^{true}(d^{true} - d^{obs})\mathbf{C}_0(d^{true})]' \\ &\quad - \mathbb{E}[(d^{true} - d^{obs})(\mathbf{B}_0(d^{true})' - \mathbf{C}_0(d^{true})')]\end{aligned} \tag{2.8}$$

where we have substituted expressions from Lemma 2 parts [5] and [6]. In general, the expression in Equation (2.8) is not zero.

Although nonzero, we can sign the terms in Equation (2.8). The second term must be negative (semi-definite).[25] In general, the first term can have arbitrary sign, but we can appeal to homophily to give it a sign. Homophily implies that $\mathbf{C}_0(d^{true})$ is positive definite for all d^{true}.[26] Since $d^{true}(d^{true} - d^{obs}) \geq 0$, this in turn implies that the expectation of the product is positive as well. In turn, the first term must be negative (semi-definite). This implies that the entire covariance

---

is measured with error, as is the case here when $d = 1$. Results can be quite different when muptiple regressors are measured with error (see Garber and Klepper, 1980). Bound, Brown and Mathiowetz (2001) derive a general formula for such cases and state the following: "[e]ven with classical assumptions, measurement error in more than one explanatory variable does not necessarily attenuate the coefficients on the variables measured with error."

[25]That is, since $d_i^{true} \geq d_i^{obs}$ for any d_i^{true} and $(\mathbf{B}_0(d^{true}) - \mathbf{C}_0(d^{true}))$ is positive definite for any d^{true} via a generalization of the Cauchy-Schwarz Inequality (see Tripathi, 1999), the expectation of their product must be positive definite.

[26]Recall that $\mathbf{C}_0(d^{true})$ is the average covariance in covariates among those who share a common friend. Within homophily in the network, we expect positive covariance in covariate space for agents who share common links.

is negative (semi-definite).

To further illustrate, we discuss a special case. When the covariate is assigned i.i.d., then $\mathbf{C}_0(d^{true}) = 0$ for all d^{true}. In this case, the expression for the covariance reduces to

$$\mathbb{E}[\frac{1}{N}(\mathbf{W}\mathbf{x})'(\mathbf{W}^* - \mathbf{W})\mathbf{x}] = -\mathbb{E}[(d^{true} - d^{obs})(\mathbf{B}_0(d^{true})'] \tag{2.9}$$

That is, even in the simplest case, there is negative correlation between the true value of the regressor, $\mathbf{W}\mathbf{x}$, and the measurement error, $(\mathbf{W}^* - \mathbf{W})\mathbf{x}$. This has important implications for the sign of the inconsistency of the observed-data estimator, and it implies that the well-known results in the "classical" model do not readily apply (even when $d = 1$).

2.3.4 True-Data Results

For purposes of completeness, before deriving the observed-data results, we provide results for the true-data estimators and thus expressions for α under each case. Lemma 4 gives the result for the linear-in-means model, while Lemma 5 gives the result for the sums model. To simplify notation, we have suppressed dependence of $\mathbf{A}_m$, etc. on degree d^{true}.

The result for the linear-in-means model is similar to the result given in Proposition 1 of Griffith (2022), but note that Assumption 5 allows for characterization in terms of expectations of $\mathbf{A}_m$ and $\mathbf{C}_m$. The linear-in-sums result, given in Lemma 5, has a similar closed form.[27]

Lemma 4. *Given Assumptions 1 - 5, if* $\mathbf{W} = \mathbf{G}$ *(linear-in-means)*

$$\alpha = \begin{bmatrix} \beta_2 \\ \beta_1\beta_2 + \beta_3 \end{bmatrix} + \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\mathbf{A}_0] \\ \mathbb{E}[\mathbf{A}_0]' & \mathbb{E}[\mathbf{F}_0] \end{bmatrix}^{-1} \left(\sum_{m=1}^{\infty} \beta_1^m \begin{bmatrix} \mathbb{E}[\mathbf{A}_m] \\ \mathbb{E}[\mathbf{F}_m] \end{bmatrix} \right) (\beta_1\beta_2 + \beta_3)$$

where $\mathbb{E}[\mathbf{F}_m] = \mathbb{E}[\mathbf{C}_m] + \mathbb{E}[\frac{1}{d^{true}}(\mathbf{B}_m - \mathbf{C}_m)]$ *for all* $m = 0, ..., \infty$ *for all* $m = 0, ..., \infty$.

---

[27] We note further that the results in Lemmas 4-5 allow for, under some conditions, deriving expressions for β_2 and $(\beta_1\beta_2 + \beta_3)$ as functions of the reduced-form parameters (α_1, α_2). In turn, this allows us to give expressions of the probability limits of the observed-data estimators in terms of these reduced-form parameters.

Lemma 5. *Given Assumptions 1 - 5, if* $\mathbf{W} = \mathbf{L}$ *(linear-in-sums)*

$$\alpha = \begin{bmatrix} \beta_2 \\ \beta_1\beta_2 + \beta_3 \end{bmatrix} + \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[d^{true}\mathbf{A}_0] \\ \mathbb{E}[d^{true}\mathbf{A}_0]' & \mathbb{E}[d^{true}\mathbf{D}_0] \end{bmatrix}^{-1} \left(\sum_{m=1}^{\infty} \beta_1^m \begin{bmatrix} \mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{true}\mathbf{D}_m] \end{bmatrix} \right) (\beta_1\beta_2 + \beta_3)$$

where $\mathbb{E}[d^{true}\mathbf{D}_m] = \mathbb{E}[(d^{true})^2\mathbf{C}_m] + \mathbb{E}[d^{true}(\mathbf{B}_m - \mathbf{C}_m)]$ *for all* $m = 0, ..., \infty$.

We further note that these results are general and apply whenever Assumptions 1 - 5 hold. That is, well-known results, such as the value of α in a "classroom" linear-in-means model (where $\mathbf{G} = \mathbf{G}^2$), are special cases of these lemmas.

2.4 The i.i.d. Case

To illustrate the main ideas and intuition behind the general results, we first show a special case, along with a sketch of its proof. For many practitioners—including most prominently when the covariate $\mathbf{x}$ is a vector of treatment indicators assigned randomly—these results will be sufficient, and one can skip the more technical general case presented in Section 2.5.[28]

2.4.1 The i.i.d. Assumption

We first state Assumption 6. This assumption imposes the restriction that covariates $\mathbf{x}$ are independent across observations. This assumption may seem strong at first, but in many field experiments, $\mathbf{x}$ is a vector of treatment statuses that is assigned randomly, and thus Assumption 6 is implied by the randomized design.[29]

Assumption 6. *(Independence of* $\mathbf{x}$*) For all* j, $\mathbf{x}_{(j)}|\mathbf{L}, \mathbf{M} \sim_{i.i.d.} \left(0, \mathbf{E_{xx}}\right)$.

Assumption 6 provides that $\mathbf{x}_{(j)}$ is independent across observations, with mean zero and constant variance. This in turn implies that the covariance of two agents' covariates is 0 *regardless of*

[28]The results presented here are special cases of the more general results in Section 2.5, as stated in Corollaries 1 (linear-in-means) model) and 2 (linear-in-sums).

[29]But see Caeyers and Fafchamps (2019) for a treatment of "exclusion bias" that may still be present in these popular designs.

those individuals' connections, in both the true and observed networks. Assumption 6 has several implications, two of which are of particular importance. First, $\mathbf{A}_0(d^{true})$, the covariance between individuals' and their friends' covariates, is 0, regardless of d^{true}. Additionally, $\mathbf{C}_0(d^{true})$, the covariance between the covariates of two agents who share a common neighbor, is also 0. The latter implies that many of the expressions in Lemma 2 simplify substantially.

We also highlight that Assumption 6 does *not* imply that $\mathbf{A}_m(d^{true})$ or $\mathbf{C}_m(d^{true})$ are 0 for $m > 0$. This is due to the fact that these values can include the covariates of the same agent in both parts of the covariance. For instance, $\mathbf{A}_1(d^{true})$ is the covariance between $\mathbf{x}_{(i)}$ and $(\mathbf{Wx})_{(j)}$ for any of j that is linked to i; if i is also linked to j—that is, if i's link to j is reciprocated by j having a link to i—then i's covariate is included as part of $(\mathbf{Wx})_{(j)}$. Similarly, $\mathbf{C}_1(d^{true})$ is the covariance between $\mathbf{x}_{(j)}$ and $\mathbf{Wx}_{(k)}$, where both j and k share a link with agent i. If agent j shares a link with k, then $\mathbf{x}_{(j)}$ is part of $\mathbf{Wx}_{(k)}$ and thus it is implausible that the covariance is 0. Importantly, this is a feature of the network structure, and it will hold even when $\mathbf{x}$ is randomly assigned.

2.4.2 Linear-in-Means

Assumption 6 greatly simplifies the setting and thus allows for simpler exposition of our main results. For the linear-in-means case, we have

$$\operatorname{plim} \hat{\alpha}^{obs} = \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{Hx}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{Hx})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{Hx})'\mathbf{Hx}] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{Hx})'\mathbf{y}] \end{bmatrix} \tag{2.10}$$

First, we apply Lemma 3 to substitute $\mathbb{E}[\frac{1}{N}(\mathbf{Hx})'\mathbf{y}] = \mathbb{E}[\frac{1}{N}(\mathbf{Gx})'\mathbf{y}]$. Next, Assumption 6 implies that $\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{Hx}] = \mathbb{E}[\mathbf{A}_0] = 0$. Making these two substitutions leads to Equation (2.11).

$$\operatorname{plim} \hat{\alpha}^{obs} = \begin{bmatrix} \mathbf{E_{xx}}^{-1} & 0 \\ 0 & \mathbb{E}[\frac{1}{N}(\mathbf{Hx})'\mathbf{Hx}]^{-1} \end{bmatrix} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{Gx})'\mathbf{y}] \end{bmatrix} \tag{2.11}$$

Next, we multiply by a particular matrix and its inverse,[30] giving Equation (2.12).

$$\begin{aligned}\operatorname{plim}\hat{\alpha}^{obs} = &\begin{bmatrix}\mathbf{E_{xx}}^{-1} & 0\\ 0 & \mathbb{E}[\frac{1}{N}(\mathbf{Hx})'\mathbf{Hx}]^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{E_{xx}} & 0\\ 0 & \mathbb{E}[\frac{1}{N}(\mathbf{Gx})'\mathbf{Gx}]\end{bmatrix}\\ &\times\underbrace{\begin{bmatrix}\mathbf{E_{xx}}^{-1} & 0\\ 0 & \mathbb{E}[\frac{1}{N}(\mathbf{Gx})'\mathbf{Gx}]^{-1}\end{bmatrix}\begin{bmatrix}\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}]\\ \mathbb{E}[\frac{1}{N}(\mathbf{Gx})'\mathbf{y}]\end{bmatrix}}_{\alpha}\end{aligned} \tag{2.12}$$

We note that the final two terms in Equation (2.12) are simply α, defined (in Definition 3) as the probability limit of the true-data estimator, which uses $\mathbf{G}$ rather than $\mathbf{H}$. Therefore,

$$\operatorname{plim}\hat{\alpha}^{obs} = \begin{bmatrix}\mathbf{I} & 0\\ 0 & \mathbb{E}[\frac{1}{N}(\mathbf{Hx})'\mathbf{Hx}]^{-1}\mathbb{E}[\frac{1}{N}(\mathbf{Gx})'\mathbf{Gx}]\end{bmatrix}\alpha \tag{2.13}$$

Next, as discussed above, Assumption 6 implies that $\mathbf{C}_0(d^{true}) = 0$ for all d^{true}. Therefore, $\mathbb{E}[\mathbf{C}_0(d^{true})] = 0$. By Lemma 2, $\mathbb{E}[\frac{1}{N}(\mathbf{Hx})'\mathbf{Hx}] = \mathbb{E}[\frac{1}{d^{obs}}\mathbf{B}_0]$ while $\mathbb{E}[\frac{1}{N}(\mathbf{Gx})'\mathbf{Gx}] = \mathbb{E}[\frac{1}{d^{true}}\mathbf{B}_0]$. Assumption 6 in turn implies that $\mathbf{B}_0$ is independent of true and observed degree: therefore $\mathbb{E}[\frac{1}{d^{obs}}\mathbf{B}_0] = \mathbb{E}[\frac{1}{d^{obs}}]\mathbb{E}[\mathbf{B}_0]$ and $\mathbb{E}[\frac{1}{d^{true}}\mathbf{B}_0] = \mathbb{E}[\frac{1}{d^{true}}]\mathbb{E}[\mathbf{B}_0]$.

Substituting these into Equation (2.13) gives the result

$$\operatorname{plim}\hat{\alpha}^{obs} = \begin{bmatrix}\mathbf{I} & 0\\ 0 & \frac{\mathbb{E}[\frac{1}{d^{true}}]}{\mathbb{E}[\frac{1}{d^{obs}}]}\mathbf{I}\end{bmatrix}\alpha \tag{2.14}$$

This result shows that, when the covariate vector $\mathbf{x}$ is i.i.d. across observations, the estimator for α_1 is consistent, while the estimator for α_2 suffers from attenuation. This is similar to the result given in Griffith (2022), but which was only proven under the assumption that $\beta_1 = 0$ and restrictions on the relationship between covariances and degree, a stronger version of Assumption 6. Our assumptions on higher-order covariances in Assumption 5 allow us to prove this result in the more

[30] As can be seen from Equation (2.12), this matrix is the covariance matrix of $[\mathbf{x}, \mathbf{Gx}]$, under the assumption of independence as stated in Assumption 6.

general case.[31]

2.4.3 Linear-in-Sums

The linear-in-sums case is quite different, due to the fact that the "measurement error" is correlated with the true value, as discussed above in Subsection 2.3.3. To see this, first note that by Slutsky's theorem,

$$\operatorname{plim} \hat{\alpha}^{obs} = \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{M}\mathbf{x}] \\ \mathbb{E}[(\frac{1}{N}\mathbf{M}\mathbf{x})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{M}\mathbf{x}] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{y}] \end{bmatrix}. \tag{2.15}$$

Next, Lemma 3 gives expressions for $\mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{y}]$ and $\mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{y}]$, which—in contrast to the linear-in-means case—in general are not the same. However, Lemma 3 provides a means to express the former as a function of the latter. Rewrite the expressions from Lemma 3 in matrix form as follows:

$$\begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{y}] \end{bmatrix} = \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=0}^{\infty} \beta_1^m \mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{obs}\mathbf{A}_0]' & \sum_{m=0}^{\infty} \beta_1^m \mathbf{D}_m^{int} \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_1\beta_2 + \beta_3 \end{bmatrix} \tag{2.16}$$

$$\begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{y}] \end{bmatrix} = \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=0}^{\infty} \beta_1^m \mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{true}\mathbf{A}_0]' & \sum_{m=0}^{\infty} \beta_1^m \mathbf{D}_m \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_1\beta_2 + \beta_3 \end{bmatrix} \tag{2.17}$$

Combining the expressions in Lines (2.16) - (2.17) then gives

$$\begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{y}] \end{bmatrix} = \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=0}^{\infty} \beta_1^m \mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{obs}\mathbf{A}_0]' & \sum_{m=0}^{\infty} \beta_1^m \mathbf{D}_m^{int} \end{bmatrix} \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=0}^{\infty} \beta_1^m \mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{true}\mathbf{A}_0]' & \sum_{m=0}^{\infty} \beta_1^m \mathbf{D}_m \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{y}] \end{bmatrix} \tag{2.18}$$

[31] We note that these results are further demonstrated by simulations in Appendix B.2.2, with both censored and random missingness. Further, we demonstrate the same patterns using real data from a randomized trial in Subsection 2.6.2.

Multiply by a matrix and its inverse to show

$$\begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{y}] \end{bmatrix} = \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{obs}\mathbf{A}_0]' & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m^{int} \end{bmatrix} \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{true}\mathbf{A}_0]' & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m \end{bmatrix}^{-1}$$
$$\begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{L}\mathbf{x}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{L}\mathbf{x}] \end{bmatrix} \underbrace{\begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{L}\mathbf{x}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{L}\mathbf{x}] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{y}] \end{bmatrix}}_{\alpha} \tag{2.19}$$

Note that the final two matrices are simply α as defined in Definition 3.

As in the linear-in-means case, Assumption 6 implies that $\mathbf{A}_0(d^{true}) = 0$ for all d^{true}. Therefore, $\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{L}\mathbf{x}] = \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{M}\mathbf{x}] = 0$. Assumption 6 also implies that $\mathbf{C}_0(d^{true}) = 0$ for all d^{true}, which in turn implies that $\mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{L}\mathbf{x}] = \mathbb{E}[d^{true}\mathbf{B}_0]$ and $\mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{M}\mathbf{x}] = \mathbb{E}[d^{obs}\mathbf{B}_0]$. The covariate $\mathbf{x}$ being i.i.d. further implies that these terms are $\mathbb{E}[d^{true}]\mathbf{E_{xx}}$ and $\mathbb{E}[d^{obs}]\mathbf{E_{xx}}$, respectively. Substitute these implications and combine Lines (2.19) and Line (2.15).

$$\begin{aligned} \operatorname{plim}\hat{\alpha}^{obs} = &\begin{bmatrix} \mathbf{E_{xx}} & 0 \\ 0 & \mathbb{E}[d^{obs}]\mathbf{E_{xx}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=1}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ 0 & \mathbf{F}^{int} \end{bmatrix} \\ &\times \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=1}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ 0 & \mathbf{F}^{true} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{E_{xx}} & 0 \\ 0 & \mathbb{E}[d^{true}]\mathbf{E_{xx}} \end{bmatrix} \alpha \end{aligned} \tag{2.20}$$

where $\mathbf{F}^{true} = \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m$ and $\mathbf{F}^{obs} = \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m^{obs}$. Next, decompose the second term on the RHS of Equation (2.20) as follows:

$$\begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=1}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ 0 & \mathbf{F}^{int} \end{bmatrix} = \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=1}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ 0 & \mathbf{F} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & (\mathbf{F}^{int} - \mathbf{F}) \end{bmatrix} \tag{2.21}$$

Next, note that

$$\begin{bmatrix} 0 & 0 \\ 0 & (\mathbf{F}^{int} - \mathbf{F}) \end{bmatrix} \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=1}^{\infty} \beta_1^m \mathbb{E}[d^{true}\mathbf{A}_m] \\ 0 & \mathbf{F}^{int} \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & (\mathbf{F}^{int} - \mathbf{F})\mathbf{F}^{-1} \end{bmatrix} \tag{2.22}$$

Substituting these into Line (2.20) then shows

$$\operatorname{plim} \hat{\alpha}^{obs} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \frac{\mathbb{E}[d^{true}]}{\mathbb{E}[d^{obs}]}\mathbf{I} \end{bmatrix} \left(\begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{I} - \mathbf{E_{xx}}^{-1}(\mathbf{F} - \mathbf{F}^{int})\mathbf{F}^{-1}\mathbf{E_{xx}} \end{bmatrix} \right) \alpha \tag{2.23}$$

which in turn simplifies to

$$\operatorname{plim} \hat{\alpha}^{obs} = \alpha + \begin{bmatrix} 0 \\ \frac{\mathbb{E}[d^{true}]}{\mathbb{E}[d^{obs}]}\mathbf{E_{xx}}^{-1}\mathbf{F}^{int}\mathbf{F}^{-1}\mathbf{E_{xx}} - \mathbf{I} \end{bmatrix} \alpha_2 \tag{2.24}$$

Equation (2.24) gives the probability limit as the sum of the true parameter α and another term. Clearly, the sign of inconsistency depends on the sign (positive/negative definiteness) of the second term.

We mention a special case here that may be important in applications. When links are missing randomly, then $\mathbb{E}[d^{obs}|d^{true}] = \frac{\mathbb{E}[d^{obs}]}{\mathbb{E}[d^{true}]}d^{true}$. That is, for any true degree, the expected *observed* degree is simply the ratio of average observed and average true degree in the population, which is a constant. By iterated expectations, for any m,

$$\begin{aligned} \mathbf{D}_m^{int} &= \mathbb{E}[d^{true}d^{obs}\mathbf{C}_m] + \mathbb{E}[d^{obs}(\mathbf{B}_m - \mathbf{C}_m)] \\ &= \mathbb{E}[d^{true}\mathbb{E}[d^{obs}|d^{true}]\mathbf{C}_m] + \mathbb{E}[\mathbb{E}[d^{obs}|d^{true}](\mathbf{B}_m - \mathbf{C}_m)] \\ &= \frac{\mathbb{E}[d^{obs}]}{\mathbb{E}[d^{true}]}\Big(\mathbb{E}[(d^{true})^2\mathbf{C}_m] + \mathbb{E}[d^{true}(\mathbf{B}_m - \mathbf{C}_m)]\Big) = \frac{\mathbb{E}[d^{obs}]}{\mathbb{E}[d^{true}]}\mathbf{D}_m \end{aligned} \tag{2.25}$$

Since $\mathbf{F}^{int} = \sum_{m=0}^{\infty} \beta_1^m \mathbf{D}_m^{int}$, this implies that $\mathbf{F}^{int} = \frac{\mathbb{E}[d^{obs}]}{\mathbb{E}[d^{true}]}\mathbf{F}$. Substitution into Equation (2.24) then shows that the second term on the right-hand side is 0. That is, when $\mathbf{x}$ is assigned randomly *and* links are missing randomly, the reduced-form linear-in-sums estimator is consistent!

We discuss other special cases in Subsection 2.5.2.[32]

2.5 Analytic Results

In this section, we provide our main analytic results. Proposition 1 gives clear results for the linear-in-means specification. For the linear-in-sums specification, Proposition 2 shows that the direction of inconsistency is, in general, theoretically ambiguous. We discuss special cases of each, including the results in Section 2.4 as corollaries of these general results.

For proofs, we refer the reader to Appendix C.1. To aid intuition, Section 2.4 gives simplfied sketches of the proofs under the additional assumption that the covariates are i.i.d. We also provide further evidence of these results using two real datasets in Section 2.6 and via simulation in Appendix B.2.

2.5.1 Linear-in-Means

First, we state the general result for the linear-in-means case, which is closely related to a result in Griffith (2022). We discuss the result in Proposition 1 through a series of remarks.

Proposition 1. *Given Assumptions 1 - 5 and* $\mathbf{W} = \mathbf{G}$ *(linear-in-means), then*

$$\operatorname{plim} \hat{\alpha}^{obs} = \alpha + \begin{bmatrix} \mathbf{Z_1}^{-1}\mathbb{E}[\mathbf{A}_0](\mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0)])^{-1} \\ -\mathbf{Z_2}^{-1} \end{bmatrix} \mathbb{E}[(\frac{1}{d^{obs}} - \frac{1}{d^{true}})(\mathbf{B}_0 - \mathbf{C}_0)]\alpha_2$$

where

- $\mathbf{Z_1} = \mathbf{E_{xx}} - \mathbb{E}[\mathbf{A_0}](\mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0)])^{-1}\mathbb{E}[\mathbf{A}_0]'$,
- $\mathbf{Z_2} = (\mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0)]) - \mathbb{E}[\mathbf{A}_0]'\mathbf{E_{xx}}^{-1}\mathbb{E}[\mathbf{A}_0]$.

[32] We note that these results are further demonstrated by simulations in Appendix B.2.2, with both censored and random missingness. Further, we demonstrate the same patterns using real data from a randomized trial in Subsection 2.6.2.

Remark 1. *The **linear-in-means** estimate of α_2 is attenuated,*[33] *regardless of the missingness pattern.*

First, Remark 1 states that $\hat{\alpha}_2^{obs}$ is in general attenuated in a particular sense. Rewrite the result in Proposition 1 as Equation (2.26).

$$\begin{aligned}\operatorname{plim} \hat{\alpha}_2^{obs} = &\left((\mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0)]) - \mathbb{E}[\mathbf{A}_0]' {\mathbf{E}_{\mathbf{xx}}}^{-1} \mathbb{E}[\mathbf{A}_0] \right)^{-1} \\ &\left((\mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{true}}(\mathbf{B}_0 - \mathbf{C}_0)]) - \mathbb{E}[\mathbf{A}_0]' {\mathbf{E}_{\mathbf{xx}}}^{-1} \mathbb{E}[\mathbf{A}_0] \right) \alpha_2. \end{aligned} \tag{2.26}$$

Note that both terms are positive definite (see Tripathi, 1999). Further, since $\frac{1}{d^{obs}} \geq \frac{1}{d^{true}}$, the first term is larger than the second, in the sense of their difference being positive definite.

Next, Remark 2 gives the sign of the inconsistency of $\hat{\alpha}_1^{obs}$. In general, it has the same sign as the product of $\mathbb{E}[\mathbf{A}_0]$ and α_2, the covariance between agents' and their neighbors' covariates and the peer effect parameter, respectively. When the network exhibits homophily on the covariates, then $\mathbb{E}[\mathbf{A}_0]$ is positive definite and $\hat{\alpha}_1^{obs}$ is inconsistent in the direction of α_2; that is, the inconsistency is the product of a positive definite matrix and α_2.[34]

Remark 2. *The sign of the inconsistency of $\hat{\alpha}_1^{obs}$ is the same as the sign of $\mathbb{E}[\mathbf{A}_0]\alpha_2$.*

Finally, we state the result when the covariates are independent across observations. As discussed above in Section 2.4, $\mathbf{x}$ being assigned i.i.d. implies that $\mathbb{E}[\mathbf{A}_0(d^{true})] = \mathbb{E}[\mathbf{C}_0(d^{true})] = 0$ for all d^{true}. Further, it also implies $\mathbf{B}_0(d^{true}) = \mathbf{E}_{\mathbf{xx}}$ for any d^{true}. This leads to Corollary 1.

Corollary 1. *Given Assumptions 1 - 5, Assumption 6, and $\mathbf{W} = \mathbf{G}$ (linear-in-means), then*

$$\operatorname{plim} \hat{\alpha}^{obs} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \frac{\mathbb{E}[\frac{1}{d_i^{true}}]}{\mathbb{E}[\frac{1}{d_i^{obs}}]} \mathbf{I} \end{bmatrix} \alpha$$

[33] We note that attenuation only strictly pertains to the case of $d = 1$, when the covariate is a scalar and thus only one regressor is measured with error. See discussion in Footnote 24.

[34] When the network is formed homophilically such that, e.g., $\mathbb{E}[\mathbf{A}_0]$ is positive definite, these results are further demonstrated by simulation in Appendix B.2.3, with results in Figure B.3.

The results in Corollary 1 are quite simple: $\hat{\alpha}_1^{obs}$ is consistent, while $\hat{\alpha}_2^{obs}$ is attenuated. These results, along with a sketch of their proof, are discusssed in more detail in Subsection 2.4.2.[35] Note that these results hold regardless of the missingness pattern and do not depend on any of hte underlying parameters.[36]

2.5.2 Linear-in-Sums

In General

Our main result for the linear-in-sums model is stated in Proposition 2. In a sharp contrast to the linear-in-means case, the direction of the inconsistency in general is ambiguous.

Proposition 2. *Given Assumptions 1 - 5 and* $\mathbf{W} = \mathbf{L}$ *(linear-in-sums),*

$$\begin{aligned}\operatorname{plim} \hat{\alpha}^{obs} = \alpha + \mathbf{P}_1^{-1}\Bigg(& \begin{bmatrix} \mathbb{E}[(d^{true} - d^{obs})\mathbf{A}_0] \\ \mathbb{E}[d^{obs}(d^{true} - d^{obs})\mathbf{C}_0] \end{bmatrix} - \sum_{m=1}^{\infty} \beta_1^m \begin{bmatrix} 0 \\ (\mathbf{D}_m - \mathbf{D}_m^{int}) \end{bmatrix} \\ & + \sum_{m=1}^{\infty} \beta_1^m \begin{bmatrix} 0 & 0 \\ \mathbb{E}[(d^{true} - d^{obs})\mathbf{A}_0]' & \sum_{m=0}^{\infty} \beta_1^m (\mathbf{D}_m - \mathbf{D}_m^{int}) \end{bmatrix} \mathbf{P}_3^{-1} \begin{bmatrix} \mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbf{D}_m \end{bmatrix} \Bigg)\alpha_2\end{aligned}$$

where

- $\mathbf{P}_1 = \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[d^{obs}\mathbf{A}_0] \\ \mathbb{E}[d^{obs}\mathbf{A}_0]' & \mathbf{D}_0^{obs} \end{bmatrix}$
- $\mathbf{P}_3 = \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=0}^{\infty} \beta_1^m \mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{true}\mathbf{A}_0'] & \sum_{m=0}^{\infty} \beta_1^m \mathbf{D}_m \end{bmatrix}$

Proposition 2 gives the probability limit as the sum of the true parameter α and a second term. The sign (definiteness) of the second term depends on underlying parameters of the model, in particular β_1. Additionally, it depends on the relationship between the covariates and the network

[35] This special case is demonstrated by simulation in Appendix B.2.2, with results in Figure B.1.

[36] In particular, the statement in Corollary 1 does not depend on $\beta_1 = 0$ as did the result in Griffith (2022). Further, the result holds regardless of the missingness mechanism.

structure. For example, the sign of the term $\mathbb{E}[(d^{true} - d^{obs})\mathbf{A}_0]$ depends on the relationship between agents' covariates and their links': if the network exhibits homophily, then this term is positive (definite). The higher-order terms $\mathbf{A}_m$ and $\mathbf{D}_m$ are even more difficult to characterize in terms of sign (positive/negative definiteness).

Without further restrictions, we have no theory-based way to make claims about the direction of inconsistency, for either the direct or peer effect. However, additional assumptions—that may be more or less plausible depending on the setting—lead to restrictions that remove the ambiguity. We discuss these below.

Restrictions on Relationship between covariates and network

Here, we give the first special case. This case, along with a sketch of its proof, is further discussed in Subsection 2.4.3 above. Corollary 2 directly follows from Proposition 2 and Assumption 6. The latter implies that $\mathbf{A}_0(d^{true}) = \mathbf{C}_0(d^{true}) = 0$ for all d^{true} and thus $\mathbb{E}[d^{true}\mathbf{A}_0] = \mathbb{E}[d^{true}\mathbf{C}_0] = 0$. We can then vastly simplify the expression given in Proposition 2.

Corollary 2. *Given Assumptions 1 - 5, Assumption 6, and* $\mathbf{W} = \mathbf{L}$ *(linear-in-sums),*

$$\operatorname{plim} \hat{\alpha}^{obs} = \alpha + \begin{bmatrix} 0 \\ \frac{\mathbb{E}[d^{true}]}{\mathbb{E}[d^{obs}]} \mathbf{E_{xx}}^{-1} \mathbf{F}^{int} \mathbf{F}^{-1} \mathbf{E_{xx}} - \mathbf{I} \end{bmatrix} \alpha_2 \tag{2.27}$$

where $\mathbf{F} = \sum_{m=0}^{\infty} \beta_1^m \mathbf{D}_m$, $\mathbf{F}^{int} = \sum_{m=0}^{\infty} \beta_1^m \mathbf{D}_m^{int}$.

Corollary 2 immediately implies that $\hat{\alpha}_1^{obs}$ is consistent. This is given in Remark 3. Note that this is the same result as in the linear-in-means case, as discussed in Remark 2. We stress that this result holds regardless of the missingness pattern, and it holds regardless of the true underlying parameters (in particular, it holds even when $\beta_1 \neq 0$).

Remark 3. *If* $\mathbf{x}$ *is i.i.d. across observations, in the linear-in-sums-case*

[1] $\hat{\alpha}_1^{obs}$ *is* ***consistent****,*

[2] The (in)consistency of $\hat{\alpha}_2^{obs}$ is theoretically ambiguous, determined by the relative magnitudes of $\mathbf{F}^{int}\mathbf{F}^{-1}$ and $\frac{\mathbb{E}[d^{obs}]}{\mathbb{E}[d^{true}]}$.

Remark 3, part [2] gives further insight into the direction of inconsistency for $\hat{\alpha}_2^{obs}$. In particular, it says that the sign is determined by the relative magnitudes of two objects. When $d = 1$ (the covariate is scalar), $\mathbf{F}^{int}\mathbf{F}^{-1}$ is a ratio. If these terms are the same, then $\hat{\alpha}_2^{obs}$ is consistent.

The assumption of i.i.d. is a strong one that is only likely to hold in special cases, most notably randomized trials. We therefore state results under a weaker condition that is more likely to hold in non-experimental settings. As defined in Assumption 7, *weak homophily* assumes that, essentially, agents that are connected in the network are more likely to have similar characteristics, in terms of their covariates $\mathbf{x}$, than agents who are not connnected.

Assumption 7. *Under **weak homophily**, $\mathbf{A}_m(d^{true})$, $\mathbf{B}_m(d^{true})$, and $\mathbf{C}_m(d^{true})$ are positive semi-definite for all $m = 0, ..., \infty$ and for all d^{true}.*

Weak homophily implies that all of the matrices in Proposition 2 are positive semi-definite. Note that the i.i.d. case as stated in Assumption 6 is a special case of weak homophily that further assumes some of these covariances are in fact 0.

Restrictions on Missingness Mechanism

Our next special case gives the result when we have restrictions on the missingness mechanism. With random missingness, links are missing randomly regardless of the degree of agents and also regardless of the covariates of the agents sharing the link. This in turn implies that for any d^{true}, $\mathbb{E}[d^{obs}|d^{true}] = pd^{true}$, where p is simply the fraction of links that are observed. In turn, this implies that $\mathbb{E}[d^{obs}\mathbf{A}_0] = p\mathbb{E}[d^{true}\mathbf{A}_0]$, etc. The assumption of random missingness vastly simplifies the expressions for the general case as stated in Proposition 2. This is shown in Corollary 3.

Corollary 3. *Given Assumptions 1 - 5, random missingness, and* $\mathbf{W} = \mathbf{L}$ *(linear-in-sums),*

$$\operatorname{plim} \hat{\alpha}^{obs} = \alpha + p(1-p) \begin{bmatrix} \mathbf{Z}_1^{-1} & 0 \\ 0 & \mathbf{Z}_2^{-1} \end{bmatrix} \Bigg(\begin{bmatrix} -\mathbb{E}[d^{true}\mathbf{A}_0](\mathbf{D}_0^{obs})^{-1} \\ \mathbf{I} \end{bmatrix} \mathbb{E}[d^{true}(d^{true}-1)\mathbf{C}_0] + \begin{bmatrix} \frac{1}{p}\mathbf{I} \\ -\mathbb{E}[d^{true}\mathbf{A}_0]'\mathbf{E_{xx}}^{-1} \end{bmatrix} \mathbb{E}[d^{true}\mathbf{A}_0] \Bigg) \alpha_2$$

where $p = \frac{\mathbb{E}[d^{obs}]}{\mathbb{E}[d^{true}]}$, $\mathbf{Z}_1 = \mathbf{E_{xx}} - \mathbb{E}[d^{obs}\mathbf{A}_0](\mathbf{D}_0^{obs})^{-1}\mathbb{E}[d^{obs}\mathbf{A}_0]'$, *and* $\mathbf{Z}_2 = \mathbf{D}_0^{obs} - \mathbb{E}[d^{obs}\mathbf{A}_0]'\mathbf{E_{xx}}^{-1}\mathbb{E}[d^{obs}\mathbf{A}_0]$.

Corollary 3 gives the probability limit of $\hat{\alpha}^{obs}$ under random missingness. This expression is notably simpler than the general case in Proposition 2. However, it remains theoretically ambiguous. To see this, note that the second term for $\hat{\alpha}_2^{obs}$ is given by

$$\operatorname{plim} \hat{\alpha}_2^{obs} = \alpha_2 + \mathbf{Z}_2^{-1} \Big(\underbrace{\mathbb{E}[d^{true}(d^{true}-1)\mathbf{C}_0]}_{\text{Term 1}} - \underbrace{\mathbb{E}[d^{true}\mathbf{A}_0]'\mathbf{E_{xx}}^{-1}\mathbb{E}[d^{true}\mathbf{A}_0]}_{\text{Term 2}} \Big) \tag{2.28}$$

Since $\mathbf{Z}_2^{-1}$ is positive definite via a generalization of the Cauchy-Schwartz Inequality (Tripathi, 1999), the sign is determined by the difference between Terms 1 and 2, as shown in Equation (2.28). In general, since $\mathbf{C}_0$ and $\mathbf{A}_0$ may take on arbitrary sign, these terms also have arbitrary sign.

The assumption of weak homophily gives us some additional insight, but it fails to remove the theoretical ambiguity. As stated in Assumption 7, weak homophily implies that both Terms 1 and 2 are positive semi-definite. Therefore, the sign of inconsistency will depend on the relative magnitudes of these two terms. Remark 4 discusses these points.

Remark 4. *In the linear-in-sums model, under random missingness, the (in)consistency of* $(\hat{\alpha}_1^{obs}, \hat{\alpha}_1^{obs})$ *is theoretically ambiguous. It remains ambiguous even under the assumption of weak homophily.*

Finally, we note the final special case that was discussed above in Subsection 2.4.3. If the covariate is assigned randomly, then it is the case that $\mathbf{C}_0$ and $\mathbf{A}_0$ are 0 for all d^{true}. Therefore, $\mathbb{E}[d^{true}(d^{true}-1)\mathbf{C}_0] = \mathbb{E}[d^{true}\mathbf{A}_0] = 0$. This in turn implies that the expression in Corollary

3 reduces to $\operatorname{plim} \hat{\alpha}^{obs} = \alpha$: the estimator is consistent. This case was discussed previously in Subsection 2.4.3 and is spelled out in Remark 5.

Remark 5. *In the linear-in-sums model, under random missingness and i.i.d. assignment,* $\hat{\alpha}^{obs}$ *is* ***consistent****.*

Regressions with number of friends

Finally, we discuss a two special case of the linear-in-sums model. In order to derive clear expressions for the results, we analyze each as a special case of random missingness. Therefore, both are special cases of Corollary 3.

First, when $\mathbf{x} = [\iota]$, where ι is a vector of ones, the linear-in-sums model reduces to the following regression:

$$y_{(i)} = \alpha_1 + \alpha_2 \underbrace{\sum_{j \neq i} L_{ij}}_{\text{No. of friends}} + \epsilon_i \tag{2.29}$$

This states a linear relationship between the number of friends and outcomes. The parameter of interest is α_2, which gives the relationship between additional friends and the outcome. Under our stated assumptions, we give the probability limit of an OLS estimate in Corollary 4.

Corollary 4. *Given Assumptions 1 - 5, random missingness,* $\mathbf{W} = \mathbf{L}$ *(linear-in-sums), and* $\mathbf{x} = [\iota]$*,*

$$\operatorname{plim} \hat{\alpha}_2^{obs} = \alpha_2 + \frac{\mathbb{V}[d^{true}] - \mathbb{E}[d^{true}]}{\frac{p}{1-p}\mathbb{V}[d^{true}] + \mathbb{E}[d^{true}]}\alpha_2$$

where $p = \frac{\mathbb{E}[d^{obs}]}{\mathbb{E}[d^{true}]}$.

Corollary 4 gives the probability limit of $\hat{\alpha}_2^{obs}$ as the sum of α_2 and a second term, which determines the direction of inconsistency. A surprising fact emerges: the direction of inconsistency is dependent on the difference between the variance and mean of the true degree distribution. That is, the direction of inconsistency is completely determined by the true degree distribution. Further,

specific network-formation models will yield specific formulas. For example, if the true network is and Erdos-Renyi random graph, d^{true} follows a Binomial distribution, and the variance is smaller than mean degree; this leads to attenuation in the estimate of α_2. Alternatively, if the true network follows a Poisson process, mean and variance are the same, yielding consistency. Thirdly, if the true network distribution is more skewed, such as generated by a preferential attachment model, $\mathbb{V}[d^{true}] > \mathbb{E}[d^{true}]$, which leads to augmentation in estimation.

We also note that the magnitude of inconsistency is determined by $\frac{p}{1-p}$, the ratio of observed to unobserved links. As $p \to 1$ (the number of links observed approaches the true number), the second term approaches 0.

Second, we give results for a related model. When $\mathbf{x} = [\iota, \mathbf{T}]$, where $\mathbf{T}$ is a $n \times 1$ vector of randomly-assigned binary treatment indicators, the linear-in-sums model can be written as Equation (2.30).

$$\mathbf{y}_i = \alpha_{1,1} + \mathbf{T}_i\alpha_{1,2} + \underbrace{\sum_{j \neq i} L_{ij}}_{\text{No. of friends}} \alpha_{2,1} + \underbrace{\sum_{j \neq i} L_{ij}\mathbf{T}_j}_{\text{No. of treated friends}} \alpha_{2,2} + \epsilon_i \tag{2.30}$$

Equation (2.30) defines a linear relationship between the outcome, one's own treatment status, number of friends, and the number of treated friends. This is a very common specification estimated in, for example, Miguel and Kremer (2004) and Oster and Thornton (2012).[37] In general, the parameters of interest are $\alpha_{1,2}$ and $\alpha_{2,2}$, the effect of one's own treatment status and the treatment status of one's network neighbors, respectively.

Corollary 5. *Given Assumptions 1 - 5, random missingness,* $\mathbf{W} = \mathbf{L}$ *(linear-in-sums), if* $\mathbf{x} =$

[37]The main specifications in Conti et al. (2013) and Lleras-Muney et al. (2023) can be viewed as versions of Equation (2.30) with additional regressors added.

$[\iota, \mathbf{T}]$, *where* $\mathbf{T}_{(i)}$ *is i.i.d. and mean zero, then*

$$\operatorname{plim} \hat{\alpha}^{obs} = \begin{bmatrix} \alpha_{1,1} \\ \alpha_{1,2} \\ \alpha_{2,1} \\ \alpha_{2,2} \end{bmatrix} + \begin{bmatrix} -\frac{\mathbb{E}[d^{true}]\left((1-p)(\mathbb{E}[(d^{true})^2]-\mathbb{E}[d^{true}])-\mathbb{E}[d^{true}]\right)}{\frac{p}{1-p}\mathbb{V}[d^{true}]+\mathbb{E}[d^{true}]} \\ 0 \\ \frac{\mathbb{V}[d^{true}]-\mathbb{E}[d^{true}]}{\frac{p}{1-p}\mathbb{V}[d^{true}]+\mathbb{E}[d^{true}]} \\ 0 \end{bmatrix} \alpha_{2,1}$$

where $p = \frac{\mathbb{E}[d^{obs}]}{\mathbb{E}[d^{true}]}$.

The result for OLS estimation of Equation (2.30) is given in Corollary 5. Importantly, estimates of the coefficients on own treatment status ($\alpha_{1,2}$) and the sum of links' treatment status ($\alpha_{2,2}$) are both consistent, while the constant term and the coefficient on number of friends are not. We also note that the coefficient on the number of friends ($\alpha_{2,1}$) is in general inconsistent, with the direction of inconsistency determined by the difference between variance and mean of the degree distribution. As discussed above, in special cases (e.g., Poisson), it is consistent.[38]

2.6 Empirical Demonstration

2.6.1 Empirical Strategy

In this section, we demonstrate the predictions of the analytic results in the prior section. We do this using two datasets. The first is the data used by Cai, de Janvry and Sadoulet (2015) in a randomized trial, while the second is the commonly-used social network data from AddHealth (Harris, 2009). We use the first to demonstrate the simpler result that holds when individuals' covariates are independent of their links' covariates (and thus Assumption 6 holds). We use AddHealth to show inconsistency when agents' and their peers' covariates are correlated, as is likely to be the case in non-randomized settings, where Assumption 7 (but not Assumption 6) may hold.

In each case, we show both the linear-in-means and linear-in-sums estimates under two dif-

[38] Since this special case has received attention in the applied literature, we demonstrate the result of Corollary 5 by simulation in Appendix B.2.4.

fernet schemes of partially-observed links. Each analysis is performed *as if* the full network data were the "true" network.[39] These missingness schemes are as follows:

[1] Under **sub-censoring**, links are observed only if the nominated person is listed among the first k links, where $k = 1, ..., 5$, under the applicable network definition ("OR" or "OUT").[40]

[2] Under **random missingness**, each "true" link is observed with probability p. That is, conditional on $L_{ij} = 1$, $M_{ij} \sim Bernoulli(p)$, where $p \in \{0.25, 0.5, 0.75, 1\}$.

We note that the "sub-censoring" startegy was performed in Griffith (2022) and also appears in Figure 1 of Sojourner (2013). In our analysis, we drop links consistent with the ordering that appears in the raw, individual data. Conditional on k, this is a deterministic process, and we accordingly present regression results for each k.

With random missingness, we omit links independently at the *edge* level. In these results, we take the original network data then randomly keep a given percent of the original links. For any $p \in (0, 1)$, this is a *random* process; accordingly, we present the mean coefficient and standard error estimates across 1000 simulations.

2.6.2 China Insurance Data

Background and Data

We demonstrate the simpler case first, using data from Cai, de Janvry and Sadoulet (2015). The experiment randomized invitations to participate in an intensive information session to learn about agricultural insurance. Agents were prompted to report up to five social links from whom they might have learned about insurance from. For purposes of estimating effects of peer exposure,

[39] In the Cai, de Janvry and Sadoulet (2015) data, up to five links were allowed per respondent. In AddHealth, up to five male and five female links were allowed per respondent (see Harris, 2009).

[40] For our results with the "OR" network definition, our sub-censored results for any rule k considers a link to have been observed if it is listed among *either* agent's first k links.

only those who were invited to later sessions were included in the main analysis. All of our results estimate the same specification that is reported in Column (2) of Table 2 of the original paper, which includes baseline controls and village fixed effects.

Since treatment was randomly assigned at the individual level, Assumption 6 is implied by the research design. Therefore, results here should be consistent with the discussion in Section 2.4 and the results in Corollaries 1 (linear-in-means) and 2 (linear-in-sums).

Sub-Censored

Table 2.1: China Insurance Sub-Censored Results
("OUT" Network)

Max Number of Links	1	2	3	4	5
Panel A: Means Definition					
$Treat_{is}$ ($\hat{\alpha}_1$)	0.028	0.028	0.029	0.029	0.030
	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)
Mean of Links' $Treat_{js}$ ($\hat{\alpha}_2$)	0.042	0.134**	0.172**	0.254***	0.291***
	(0.034)	(0.052)	(0.068)	(0.076)	(0.082)
R-squared	0.108	0.112	0.113	0.117	0.119
Panel B: Sums Definition					
$Treat_{is}$ ($\hat{\alpha}_1$)	0.029	0.027	0.028	0.030	0.030
	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)
Sum of Links' $Treat_{js}$ ($\hat{\alpha}_2$)	-0.009	0.054*	0.060***	0.051***	0.058***
	(0.040)	(0.027)	(0.022)	(0.018)	(0.017)
R-squared	0.107	0.110	0.113	0.114	0.119

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. Standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1.

Sub-censored results are shown in Table 2.1.[41] In Panel A, for the linear-in-means model, we see that the peer effects estimates ($\hat{\alpha}_2$) are attenuated (smaller in magnitude) when only a subset

[41]These results use the "OUT" link definition as do the original authors: note that the results in Panel A, Column 5 are the same as those in Table 2, Column (2) of Cai, de Janvry and Sadoulet (2015). Results are qualitatively similar if we instead employ the "OR" link definition. See Appendix Table B.1.

of links are observed. Further, looking across the columns of Panel A, we see that there is more attenuation when fewer links are observed. This is consistent with the result in Corollary 1, which gives a formula that implies that fewer observed links implies more attenuation. We also note that the estimates for the effect of own treatment status ($\hat{\alpha}_1$) are unaffected by censoring, consistent with the discussion in Section 2.4 and Corollary 1.

Panel B gives results for the linear-in-sums specification. Here, we see results that are consistent with Corollary 2. In this setting, there is very little variance in "true" reported degree (defined as all links observed) since most individuals reported the maximum of 5 links. As discussed in Remark 3, $\hat{\alpha}_1^{obs}$ is unaffected by censoring. Interestingly, the estimate of $\hat{\alpha}_2^{obs}$ is also relatively unaffected, which suggests that for each k, the two terms that determine inconsistency have the same magnitude.

Randomly Missing Links

Results with randomly missing links are shown in Table 2.2.[42] The linear-in-means results are given in Panel A. These results are qualitatively similar to the results in Table 2.1. First, the estimated direct effect ($\hat{\alpha}_1$) is unaffected by missing links. Second, estimated peer effect ($\hat{\alpha}_2$) is attenuated when only a subset of links are observed. Further, the fewer links that are observed, the more extreme the attenuation is.[43]

Table 2.2, Panel B gives the results for the linear-in-sums specification with randomly missing links. There, we see patterns that are consistent with the analytic results above. First, the estimated effect of own treatment status ($\hat{\alpha}_1$) is unaffected by missing links, which follows from Corollary 2. Second, the estimated peer effect ($\hat{\alpha}_2$) is also unaffected by the fact that links are missing. This is the result discussed in Remark 5 and that emerges as a special case of Corollaries 2 and 3.

---

[42] We also perform this exercise using an "OR" network definition, with qualitatively similar results as shown in Appendix Table B.2.

[43] We note that the number of observations in Panel A varies across simulations due to the fact that, in each simulation, we estimate the regression conditional on the mean being defined, which requires agents to have at least one link. If all of an agents' links are unobserved in a given simulation run, we omit that observation from the subsequent regression.

Table 2.2: China Insurance with Randomly Missing Links
("OUT" Network)

Percent of Links Observed	25%	50%	75%	100%
Panel A: Means Definition				
$Treat_{is}$ ($\hat{\alpha}_1$)	0.029	0.030	0.031	0.033
	(0.036)	(0.033)	(0.032)	(0.032)
	[0.016]	[0.006]	[0.003]	[0.000]
Mean of Links' $Treat_{js}$ ($\hat{\alpha}_2$)	0.083	0.116	0.186	0.290
	(0.052)	(0.056)	(0.069)	(0.082)
	[0.046]	[0.040]	[0.038]	[0.000]
N	963.987	1234.553	1284.207	1296
Panel B: Sums Definition				
$Treat_{is}$ ($\hat{\alpha}_1$)	0.031	0.031	0.032	0.033
	(0.032)	(0.032)	(0.032)	(0.032)
	[0.002]	[0.003]	[0.003]	[0.000]
Sum of Links' $Treat_{js}$ ($\hat{\alpha}_2$)	0.061	0.059	0.058	0.058
	(0.034)	(0.024)	(0.019)	(0.017)
	[0.030]	[0.016]	[0.009]	[0.000]
N	1296	1296	1296	1296

Notes: Based on 1,000 simulations, with links kept with indicated probability. 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. Average standard errors (across simulations) in parentheses, clustered by village. Standard deviation of point estimates (across simulations) in brackets.

2.6.3 AddHealth

In this subsection, we present simulations using network data and outcomes from AddHealth. We simulate random and censored missingness, and we show that these results are consistent with the analytic results discussed above.

Background

The National Longitudinal Study of Adolescent and Adult Health (AddHealth) is a long-running panel survey (Harris, 2009). With both academic and behavioral outcomes as well as detailed social network data, AddHealth has been used to study associations between peers and

outcomes in a large number of studies. For our empirical exercises, we estimate a form of the regression given in Equation (2.31).

$$y_{is} = \mathbf{X}_{is}\alpha_1 + \big(\sum_{j \neq i} W_{ij}\mathbf{X}_{js}\big)\alpha_2 + \gamma_s + \epsilon_{is} \tag{2.31}$$

where y_{is} is some outcome for student i in school s and γ_s is a school fixed effect. The coefficients of interest are the vectors α_1 and α_2, which give the (conditional) association between the outcome y_{is} and the individual's own and peers' characteristics, respectively.

In all specifications, $\mathbf{X}_{is}$ is a vector of ten covariates, which are summarized in Appendix Table B.3, Panel A. In contrast to the China insurance data, these regressors are not randomly assigned, and thus these results demonstrate the predicted patterns under weak homophily as defined in Assumption 7. To preserve space, we only present coefficients for Age, Grade, and Female. We focus on the outcome GPA in All Subjects, but we estimate our results for a total of six outcomes, which are summarized in Appendix Table B.3, Panel B. All show similar patterns.

Sub-censored Results

Sub-censored results are presented in Table 2.3.[44] For the means specification in Panel A, the results show attenuation for estimates of peer characteristics. When fewer links are observed (as indicated by k), estimates are *more* attenuated. This conforms to analytic predictions above.

The coefficients on own characteristics further show results that are remarkably consistent with the analytic results. If we assume that individuals' and their peers' covariates are positively correlated—as is evident in the data (unreported)—we expect the inconsistency to be in the direction as the corresponding α_2. The coefficients on $\overline{Age}$ and $\overline{Female}$ are both negative, and therefore the results skew more negative with more censoring (smaller k). The coefficient on $\overline{Grade}$, in contrast, is positive, and therefore lower k leads to inconsistency in the positive direction.

The results for the linear-in-sums specification in Table 2.3, Panel B, are quite different. The

[44]Note that we use the "OR" network definition for all results in the main text. Results using the asymmetric "OUT" definition are in Appendix Tables B.4 and B.5.

estimated coefficients on peer characteristics appear to be *augmented* for all three covariates; that is, they are larger when fewer links are observed. This is the exact opposite pattern to the linear-in-means results shown in Panel A. The coefficients for the three covariates are inconsistently negative, positive, and negative. This is consistent with the theoretical result given in Proposition 2.

We further note that the patterns shown in Table 2.3 for the outcome "GPA in All Subjects," for both means and sums specifications, are also observed across many other outcomes. We graph the coefficients on the age covariate for six different outcomes in Figure 2.2. Coefficients for grade and female, showing similar patterns, are in Appendix Figures B.6 and B.7, respectively.

Random Missingness Results

The AddHealth results for random missingness are shown in Table 2.4. These results are based on 1000 simulations for each $p \in \{0.25, 0.5, 0.75\}$. The results in Panel A are qualitatively identical to the sub-censored results in Table 2.3, Panel A. The peer effect coefficients $(\hat{\alpha}_2)$ are attenuated when fewer than 100% of links are observed, and the coefficients are approaching 0 as the percent observed gets smaller. The coefficients on own characteristics $(\hat{\alpha}_1)$ are all inconsistent in the direction of the corresponding peer effect coefficients $(\hat{\alpha}_2)$.

For the linear-in-sums specification, with random missingness, the sign of inconsistency is theoretically ambiguous, as discussed in Remark 4. However, our results in Table 2.4, Panel B show clear patterns. With fewer links observed, coefficients on Age, Grade, and Female, are inconsistently negative, positive, and negative, matching the signs of the corresponding peer effect coefficients. Estimates of coefficients of peer characteristics are augmented: the fewer links that are observed, the larger those estimates become in magnitude.[45]

[45]To show that these patterns are robust to other outcomes, in Figure 2.3 we graph estimated coefficient for Age with randomly missing links. Results for Grade and Female are similarly consistent across outcomes, as shown by Appendix Figures B.8 and B.9.

2.7 Conclusion

In this paper, we analyze the (in)consistency of reduced-form peer effects estimators when reserachers only observe a subset of the true links in the network. We show that the results depend crucially on the model being estimated. Linear-in-means estimates are a special case of "classical" measurement error, while linear-in-sums estimates may be inconsistent in theoretically-ambiguous directions. We discuss special cases of the latter, including cases where the estimator is consistent even with mismeasured networks. We then demonstrate these results with two datasets, including one where the covariate was assigned randomly as part of a field experiment. These results are further confirmed by a simulation study in Appendix B.2.

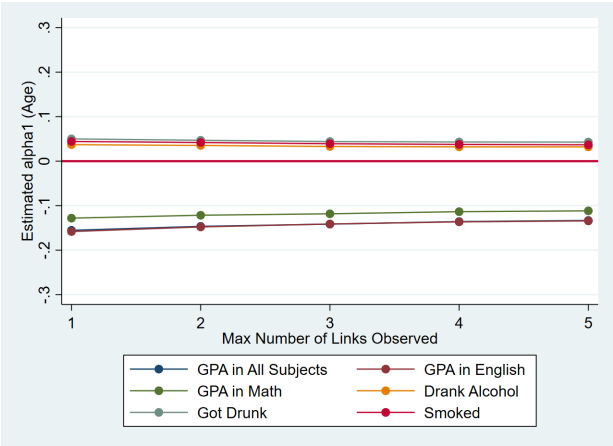
The implications of mismeasured networks is an active research area. Our contrasting results for the linear-in-means and linear-in-sums models suggest that care should be taken in extrapolating results beyond the estimator under study, and that more work is needed to provide guidance to applied researchers. In this regard, Lewbel, Qu and Tang (2022), Cai (2022), and Hsieh et al. (2024) are promising steps.

Table 2.3: AddHealth Sub-censored Results
("OR" Network)

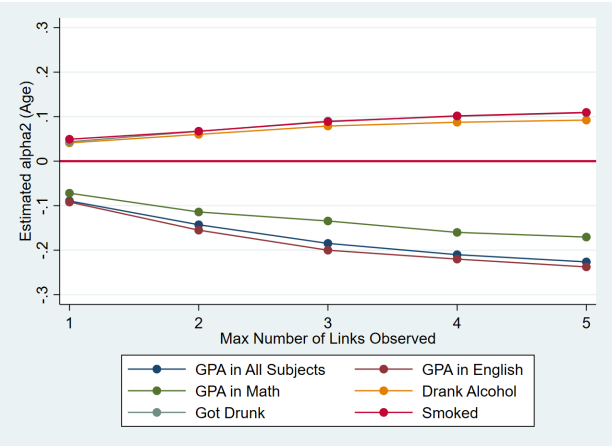
Censoring Rule (k)	1	2	3	4	5
Panel A: Means Specification					
Coefficients on Own Characteristics ($\hat{\alpha}_1$)					
Age	-0.155***	-0.146***	-0.141***	-0.136***	-0.133***
	(0.012)	(0.011)	(0.011)	(0.011)	(0.011)
Grade	0.167***	0.155***	0.146***	0.145***	0.141***
	(0.014)	(0.015)	(0.016)	(0.016)	(0.016)
Female	0.153***	0.157***	0.157***	0.159***	0.160***
	(0.011)	(0.012)	(0.012)	(0.012)	(0.012)
Coefficients on Peer Characteristics ($\hat{\alpha}_2$)					
Age	-0.089***	-0.143***	-0.185***	-0.210***	-0.226***
	(0.011)	(0.017)	(0.020)	(0.024)	(0.025)
Grade	0.080***	0.136***	0.181***	0.200***	0.218***
	(0.012)	(0.017)	(0.020)	(0.024)	(0.026)
Female	-0.028**	-0.039**	-0.044**	-0.048**	-0.054***
	(0.011)	(0.015)	(0.018)	(0.018)	(0.019)
Panel B: Sums Specification					
Coefficients on Own Characteristics ($\hat{\alpha}_1$)					
Age	-0.159***	-0.152***	-0.148***	-0.145***	-0.143***
	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)
Grade	0.174***	0.171***	0.170***	0.167***	0.165***
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Female	0.151***	0.154***	0.156***	0.156***	0.156***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Coefficients on Peer Characteristics ($\hat{\alpha}_2$)					
Age	-0.027***	-0.024***	-0.021***	-0.019***	-0.018***
	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)
Grade	0.020***	0.018***	0.017***	0.016***	0.015***
	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
Female	-0.013*	-0.008*	-0.006*	-0.005	-0.005*
	(0.007)	(0.005)	(0.004)	(0.003)	(0.003)

Dependent Variable: GPA in All Subjects. N = 32,156 in all specifications. 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. *** p<0.01, ** p<0.05, * p<0.1.

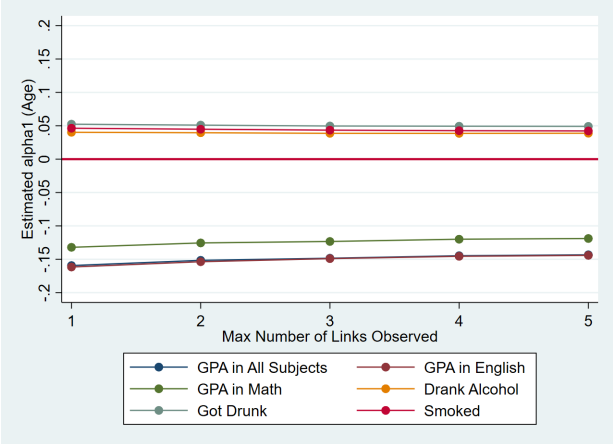
Figure 2.2: Sub-censored AddHealth Estimates
("OR" Network, Age)



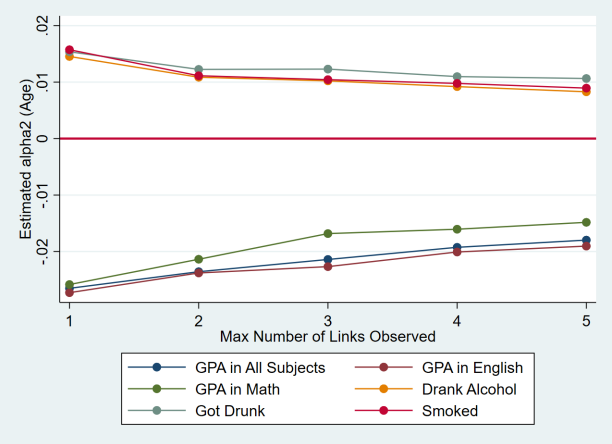
a Coefficients for Age_{is}
(Means Specification)



b Coefficients for $\overline{Age}_{is}$
(Means Specification)



c Coefficients for Age_{is}
(Sums Specification)



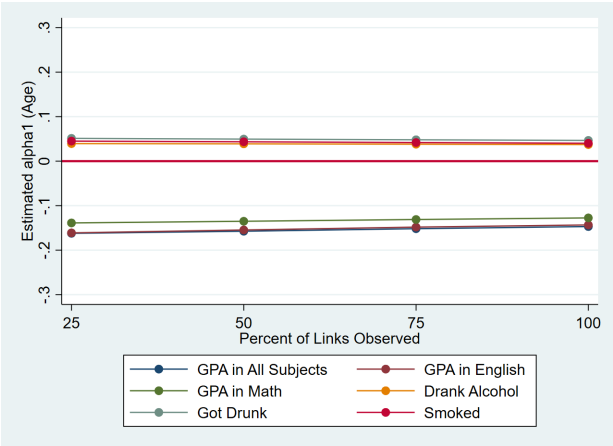
d Coefficients for $\overline{Age}_{is}$
(Sums Specification)

Table 2.4: AddHealth Results with Random Missingness ("OR" Network)

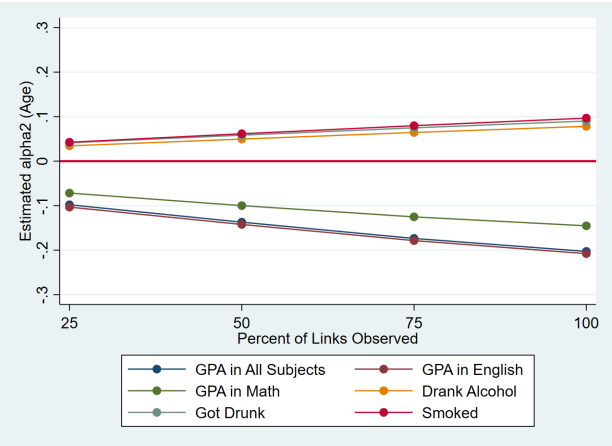
Percent Observed	25%	50%	75%	100%
Panel A: Means Specification				
Coefficients on Own Characteristics ($\hat{\alpha}_1$)				
Age	-0.162	-0.157	-0.152	-0.147
	0.012	0.011	0.011	0.010
Grade	0.172	0.166	0.162	0.160
	0.015	0.015	0.015	0.015
Female	0.152	0.153	0.154	0.154
	0.011	0.010	0.010	0.010
Coefficients on Peer Characteristics ($\hat{\alpha}_2$)				
Age	-0.099	-0.138	-0.172	-0.203
	0.011	0.014	0.018	0.020
Grade	0.093	0.131	0.163	0.190
	0.013	0.015	0.018	0.021
Female	-0.026	-0.033	-0.037	-0.042
	0.011	0.012	0.014	0.016
Observations	31,130.840	38,059.430	40,317.510	41,286
Panel B: Sums Specification				
Coefficients on Own Characteristics ($\hat{\alpha}_1$)				
Age	-0.174	-0.170	-0.166	-0.164
	0.011	0.011	0.011	0.011
Grade	0.187	0.185	0.183	0.182
	0.012	0.012	0.012	0.012
Female	0.154	0.155	0.156	0.156
	0.009	0.009	0.009	0.009
Coefficients on Peer Characteristics ($\hat{\alpha}_2$)				
Age	-0.007	-0.006	-0.006	-0.006
	0.002	0.001	0.001	0.001
Grade	0.005	0.004	0.004	0.004
	0.002	0.002	0.001	0.001
Female	-0.002	-0.001	0.000	0.001
	0.006	0.004	0.003	0.003
Observations	43,650	43,650	43,650	43,650

Notes: Average point estimates across 1000 simulations. Average standard errors, clustered by school, across simulations in parentheses. Number of observations is average across simulatoins. Within simulations, individuals who have no links for a given draw are omitted in Panel A (but not in Panel B).

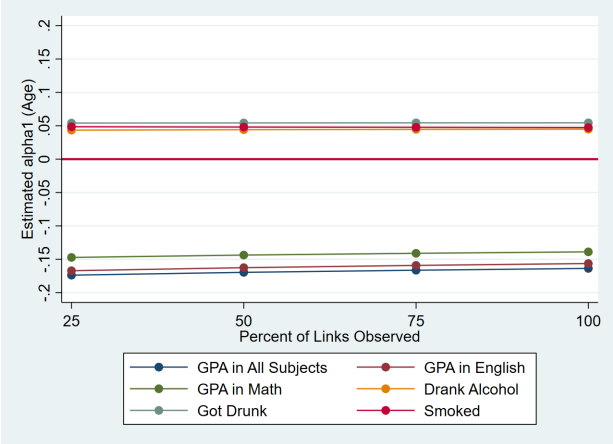
Figure 2.3: AddHealth Estimates with Random Missingness
("OR" Network, Age)



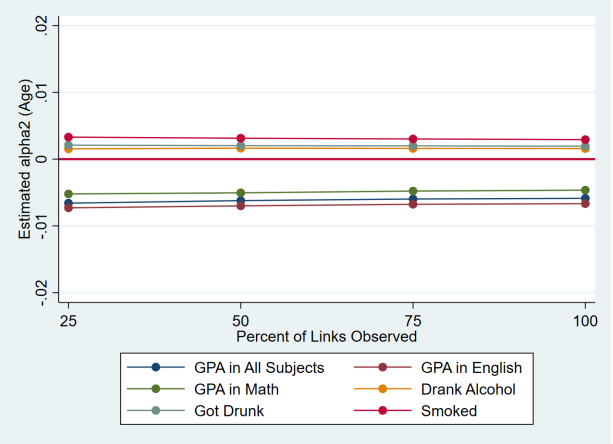
a Coefficients for Age_{is}
(Means Specification)



b Coefficients for $\overline{Age}_{is}$
(Means Specification)



c Coefficients for Age_{is}
(Sums Specification)



d Coefficients for $\overline{Age}_{is}$
(Sums Specification)

Chapter 3

EFFICIENT STRATIFIED SAMPLING FOR DIFFUSIONS IN NETWORKS: A FEASIBLE NEYMAN ALLOCATION APPROACH

Here, we introduce a novel approach to sampling in network diffusion processes, with a focus on epidemiological models. We present a feasible version of the Neyman allocation that optimizes sampling across population strata based on the network degree distribution. Our method requires knowledge of only the first moments of the degree-based strata, not the full network structure. We develop the theoretical foundations for optimal stratified sampling in this context and propose a feasible approximation of the within-strata variance based on epidemiological modeling. Through simulations on randomly generated networks, we demonstrate that our proposed methods show efficiency gains over simple random sampling, particularly in the early stages of diffusion processes. The feasible method closely approximates the performance of the optimal method while being implementable in practice. Our findings can inform sampling strategies for monitoring real-world epidemics with limited resources.

3.1 Introduction

The spread of information, behaviors, and diseases through networks is a critical area of study across multiple disciplines, from epidemiology to sociology and economics. Accurately measuring these diffusion processes is crucial for developing effective interventions and policies. However, collecting data frequently for a large population can be prohibitively expensive and time consuming.[1] This challenge necessitates the development of efficient sampling strategies using limited resources, especially during for a time such as the recent COVID-19 pandemic.

This paper introduces a novel approach to sampling in diffusion processes on networks, with

[1] Clark and Turner (2021) discusses the difficulty of monitoring epidemics through Surveys.

a particular focus on epidemiological models. We present a feasible version of the Neyman allocation that optimizes the allocation of sampling resources across different strata of the population, based on the degree distribution of the population network.[2] Our method builds upon classical stratified sampling techniques and utilizes the dynamic nature of diffusion processes in networks to construct a feasible method. We emphasize that this method requires no knowledge of the entire underlying network structure, but only the first moments of the strata based on the degree distribution.

We begin by outlining a simple diffusion model based on the well-known Susceptible-Infectious-Recovered (SIR) framework, where our interest is to estimate the proportion of agents in each state (S or I or R) for each sampling period. We develop the theoretical foundations for optimal stratified sampling in this context, demonstrating how the optimal Neyman allocation depends on the within-strata variance. Given the oracle property of the method and thus the practical limitations, we propose a feasible approximation of the variance based on epidemiology modeling with network structure.

The current literature on a feasible version is known as the adaptive Neyman allocation strategy, where the earlier stage of the samples are used to construct the required variance for the allocation.[3] Our work contrasts this concept as we can make use of the theoretical foundation of network based epidemiology modeling.[4] This is an advantageous method to save resources since there is not only no need to run a pilot run but also more efficient as we have prior information on how the variance would evolve over time through the epidemic models.[5]

To evaluate the performance of our proposed method, we conduct simulations on randomly generated networks based on a log-normal degree distribution. We report the optimal sampling method (Neyman allocation) along with the feasible method and the traditional stratified random sampling method. We compare it to simple random sampling and demonstrate that our proposed

---

[2]Based on the seminal work by Neyman (1934).

[3]See recent works of Zhao (2023); Dai, Gradu and Harshaw (2023), Blackwell, Pashley and Valentino (2023).

[4]See Paré, Beck and Başar (2020) for recent development on the analysis of epidemics over networks.

[5]Cai and Rafi (2024) point out the caveats of small pilot runs for Neyman allocation.

methods show gain in efficiency, particularly in the early stages of the diffusion process.

This paper proceeds as follows. In Section 2, we describe the a SIR diffusion model. In Section 3, we obtain the optimal sampling scheme, namely the Neyman allocation. Section 4 provides the feasible sampling strategy through a network diffusion framework. In Section 5, we show the performance of each method through simulation studies and Section 6 concludes.

3.2 A Simple Diffusion Model

We define the diffusion process under the context of epidemiology, by modeling the phenomenon of pathogen (causing a disease) spread as a state change of agents (nodes) throughout time. Following Barabási (2016), an agent can be in one of the three states:

- Susceptible (S): Healthy individuals who have not yet contacted the pathogen.
- Infectious (I): Contagious individuals who have contacted the pathogen and hence can infect others.
- Recovered (R): Individuals who have been infected before, but have recovered from the disease, hence are not infectious.

These states can form the two of the most frequently used models, the so-called SI and SIR model. With slightly abusing the notations, we denote the total number of agents in each state as $S(t), I(t), R(t)$, and the corresponding proportion among the population as $s(t), i(t), r(t)$ at time t.

For a diffusion process, our objective is to estimate the proportion of agents in each state at each time t. Formally, we estimate $\mathbb{E}[x_j^\tau(t)]$, the mean of a binary valued random variable for each agent j of each state $\tau \in \{S, I, R\}$. For example, in the case of agents in the infectious state ($\tau = I$), we estimate $\mathbb{E}[x_j^I(t)] = i(t)$ where

$$x_j^I(t) = \begin{cases} 1 & \text{if agent } j \text{ is infected} \\ 0 & \text{otherwise.} \end{cases}$$

Assume that $\mathbb{E}\left[x_j^2(t)\right] < \infty$ for all t.

3.3 Stratified Sampling

For each epidemic or diffusion process, agents are allocated in to $N_V \geq 1$ disjoint "bins" or "strata" (e.g. defined by degree). For a given set of bins, let q_v be the (known) probability of being in bin v in the population (so, $\sum_v q_v = 1$). Note that all variables in this section are time varying and equivalently applicable for all τ but we drop the notations for simplicity.

3.3.1 Optimal Stratified Sampling

For a fixed sample size N, let N_v be the number of agents in each stratum v where $\sum_v N_v = N$. So, the estimator of μ is

$$\hat{\mu}(\mathbf{N}) = \sum_v q_v \left(\frac{1}{N_v} \sum_{j=1}^{N_v} x_{jv} \right) \tag{3.1}$$

where $\mathbf{N} = (N_1, ..., N_V)$. Since $\mathbb{E}\left[x_{jv}\right] = \mu_v$ for all v, $\hat{\mu}(\mathbf{N})$ is always unbiased. The variance is given by

$$\mathbb{E}\left[(\hat{\mu}(\mathbf{N}) - \mu)^2\right] = \mathbb{E}\left[\left(\sum_v q_v \left(\frac{1}{N_v} \sum_{j=1}^{N_v} x_{jv} \right) - \mu \right)^2 \right]$$

The solution that minimizes the variance is N_v^* for all v as follows:

$$N_v^* = \frac{q_v \sigma_{\epsilon,v}}{\sum_w q_w \sigma_{\epsilon,w}} N.$$

where $N_v^* \geq 1$ for all v for $\hat{\mu}\left(\mathbf{N}^*\right)$ to be unbiased. Note that $\sigma_{\epsilon,v}$ is the within strata standard deviation of stratum v. This sampling scheme is also known as the Neyman Allocation. Therefore, the optimal allocation depends only on the within-strata variance. Further, the variance at $\mathbf{N}^*$ is given by:

$$\mathbb{V}\left[\hat{\mu}\left(\mathbf{N}^*\right)\right] = \frac{1}{N} \left(\sum_v q_v \sigma_{\epsilon,v} \right)^2 + \sum_v \sum_w q_v q_w \rho_{vw} \sigma_v \sigma_w$$

where ρ_{vw} is the expected correlation of x_{jv} and x_{jw}.

3.3.2 Simple and Stratified Random Sampling

First, we compare simple random sampling, where N are drawn at random from the population. Define this estimator as $\hat{\mu}_{SRS}$ and the variance is

$$\mathbb{V}\left[\hat{\mu}_{SRS}\right] = \frac{1}{N}\sum_{v} q_v\left(\sigma^2_{\alpha,v} + \sigma^2_{\epsilon,v}\right) + \frac{N-1}{N}\sum_{v}\sum_{w} q_v q_w \rho_{vw}\sigma_v\sigma_w$$

Next, define stratified random sampling, where $q_v = \frac{N_v}{N}$ for all v. That is, the sample is divided up according to the proportions of agents in each stratum in the population. Let $\mathbf{q}$ be the vector of frequencies in each stratum in the population. Therefore $\mathbf{N} = N\mathbf{q}$. So,

$$\mathbb{V}\left[\hat{\mu}\left(N\mathbf{q}\right)\right] = \mathbb{E}\left[\left(\hat{\mu}\left(N\mathbf{q}\right) - \mu\right)^2\right] = \frac{1}{N}\sum_{v} q_v\sigma^2_{\epsilon,v} + \sum_{v}\sum_{w} q_v q_w \rho_{vw}\sigma_v\sigma_w$$

3.3.3 Variance Comparisons

The gains in variance moving from SRS to unbiased stratified random sampling are given as

$$\mathbb{V}\left[\hat{\mu}_{SRS}\right] - \mathbb{V}\left[\hat{\mu}\left(N\mathbf{q}\right)\right] = \frac{1}{N}\left(\sum_{v} q_v\sigma^2_{\alpha,v} - \sum_{v}\sum_{w} q_v q_w \rho_{vw}\sigma_v\sigma_w\right)$$

which must always be (weakly) positive due to the Cauchy Schwartz Inequality. Next,

$$\mathbb{V}\left[\hat{\mu}\left(N\mathbf{q}\right)\right] - \mathbb{V}\left[\hat{\mu}\left(\mathbf{N}^*\right)\right] = \frac{1}{N}\left(\mathbb{E}_v\left[\sigma^2_{\epsilon,v}\right] - \mathbb{E}_v\left[\sigma_{\epsilon,v}\right]^2\right) = \frac{1}{N}\mathbb{V}_v\left[\sigma_{\epsilon,v}\right]$$

That is, the gain for the optimal stratified sampling is determined by the variance in the unobserved variance $\mathbb{V}_v\left[\sigma_{\epsilon,v}\right]$. In the corner case where this is zero, there is no gain.

3.4 Application

In section 3.3, we have seen that, regardless of how agents are categorized, the optimal sampling strategy is to sample

$$N_v^{\tau *}(t) = \frac{q_v \sigma_{\epsilon,v}^{\tau}(t)}{\sum_w q_w \sigma_{\epsilon,w}^{\tau}(t)} N \tag{3.2}$$

number of samples from each group v for each time t. If the true standard deviation $\sigma_{\epsilon,v}^{\tau}(t)$ is known, we can simply obtain the *almost* optimal estimator, in terms of sampling variance.[6] For a real world application, one can expect to estimate the standard deviation and replace it with the true value instead.

However, for each independent diffusion process, we cannot estimate the standard deviation until we observe the data, which is the estimator of interest itself. For example, in a SI model, we wish to estimate the proportion of infected agents at time t, $\mathbb{E}\left[x_j^I(t)\right] = \sum_v q_v \mathbb{E}\left[x_{j,v}^I(t)\right]$. Since the elements in $x_{j,v}^I(t)$ takes binary values of 0 and 1, the standard deviation is simply

$$\sigma_{\epsilon,v}^I(t) = \sqrt{i_v(t)\left[1 - i_v(t)\right]}$$

where, with an abuse of notation, $i_v(t)$ is the proportion of infected agents in group v. Then, it is obvious that we cannot replace $i_v(t)$ with $\hat{i}_v(t)$ by sampling $N_v^*(t)$ agents instead of Nq_v since $N_v^*(t)$ is dependent on $\sigma_{\epsilon,v}^I(t)$.

In order to overcome this, instead of estimating the standard deviation, we take an approach to directly approximate $i_v^{\tau}(t)$. In other words, we rely on the dynamics of the diffusion processes – change of the proportion of each state.

---

[6]The estimator is *almost* optimal due to the assumption of $N_v^*(t) \geq 1$ for the unbiasness of the estimator. Even when there is zero variance within a strata, we must sample at least one from each.

3.4.1 Diffusion on a Network

To explore the dynamics of the diffusion process given the network structure of agents, we define an undirected network $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with the set agents $\nu_i \in \mathcal{V}$ and edge set $\mathcal{E}$ where $|\mathcal{V}| = \mathcal{N}$. We denote the number of agents with degree k as $\mathcal{N}_k$ and the corresponding probability mass of each degree as $p_k = \frac{\mathcal{N}_k}{\mathcal{N}}$. Let us additionally denote $S_k(t), I_k(t), R_k(t)$ as the total number of agents with degree k in each state of S, I, R at time t. We finally denote $s_k(t), i_k(t), r_k(t)$ as the proportion of agents in each state of each degree, i.e. $i_k(t) = \frac{I_k(t)}{\mathcal{N}_s}$ denotes the proportion of infected agents among those with degree k and note that

$$\frac{I(t)}{\mathcal{N}} = \sum_k p_k i_k(t)$$

is the total proportion of infected agents among the population.

The dynamics of the diffusion process for agents with degree k start with the change of proportion of infected agents, namely,

$$\frac{di_k(t)}{dt} = \beta k s_k(t) \Theta_k(t) - \gamma i_k(t)$$

where β is the *rate of transmission* and γ is *the rate of recovery*. We define

$$\Theta_k(t) = \frac{1}{k} \sum_{k'} (k' - 1) p_{k'} i_{k'}(t)$$

as the fraction of infected nodes in the neighborhood of a susceptible node with degree k.[7] Given

---

[7]For an agent with degree k, the probability that a link points to another agent with degree k' is

$$\frac{k' p_{k'}}{k}.$$

At least one link of each infected agent is connected to another infected agent, the one that transmitted the infection. Thus, the proportion of links connected to agents of degree k' who are available for future transmission is

$$\frac{(k' - 1) p_{k'}}{k}$$

$s_k(t) + i_k(t) + r_k(t) = 1$, we derive the dynamics of $\Theta_k(t)$ as

$$\begin{aligned}\frac{d\Theta_k(t)}{dt} &= \frac{1}{k}\sum_{k'}(k'-1)p_{k'}\frac{di_{k'}(t)}{dt} \\ &= \frac{\beta}{k}\sum_{k'}\left[(k')^2 - k'\right]\left[(1 - i_{k'}(t) - r_{k'}(t))\, p_{k'}\Theta_{k'}(t) - \gamma i'_k(t)\right].\end{aligned}$$

By the definition of γ, we have

$$\begin{aligned}\frac{dr_k(t)}{dt} &= \gamma i_k(t) \\ s_k(t) &= 1 - i_k(t) - r_k(t)\end{aligned}$$

as designed by the model, which shows that the dynamics of $i_k(t)$ and $\Theta_k(t)$ derives the entire diffusion process. The dynamic system can be defined as a set of joint differential equations of

$$\frac{di_k(t)}{dt} = f_i\Big(i_k(t), \Theta_k(t)\Big) \tag{3.3}$$

$$\frac{d\Theta_k(t)}{dt} = f_\Theta\Big(i_k(t), \Theta_k(t)\Big) \tag{3.4}$$

for all k.

3.4.2 A Feasible Sampling Strategy

The dynamic system above allows us to approximate a discrete time values of the proportion of each state. We first approximate $\tilde{s}_k(t)$, $\tilde{i}_k(t)$, and $\tilde{r}_k(t)$ via the Runge-Kutta method.[8] Then,

---

and thus the proportion of infected agents of all possible degrees in the neighborhood of the agent with degree k is

$$\Theta_k(t) = \sum_{k'}\frac{(k'-1)p_{k'}i_{k'}(t)}{k}.$$

[8]We first approximate $\tilde{i}_k(t)$ as

$$\tilde{i}_k(t+1) = \tilde{i}_k(t) + \frac{1}{6}(\phi_1 + 2\phi_2 + 2\phi_3 + \phi_4)$$

we calculate the approximated standard deviation $\tilde{\sigma}^{\tau}_{\epsilon,k}(t)$ for each τ, and thus obtain a stratified sampling strategy of

$$\tilde{N}^{\tau}_{k}(t) = \frac{q_k \tilde{\sigma}^{\tau}_{\epsilon,k}(t)}{\sum_w q_w \tilde{\sigma}^{\tau}_{\epsilon,w}(t)} N \tag{3.5}$$

where each strata is naturally defined as a group of agents with the same degree, e.g. k.

Despite its usability, whether the strategy is feasible in a realistic scenario is questionable. There are two main caveats. First, obtaining the degree information of all agents in the population is extremely costly as it would require to sample the entire population. Second, even if the information is obtainable, it is possible that there would be only one agent with a certain value of degree in the population. This possibility, along with the assumption of $\tilde{N}^{\tau *}_{k}(t) \geq 1$ for the unbiasness of the estimators, results in sampling the one agent for every sample period which is not an optimal choice since the variance would always be zero in a strata with one agent, i.e. the optimal numbers to sample would be zero as well.

In order to overcome these shortcomings, we coarsen the strata with respect to the degree of agents so that each strata has enough agents to sample from. A simple example configuration would be to reallocate agents into three groups – *high*, *medium*, and *low* degree – and assume all agents within the same strata to have the same degree (i.e. average degree). Let k_h, k_m, and k_l

---

where

$$\begin{aligned}
\phi_1 &= f_i\left(\tilde{i}_k(t),\ \tilde{\Theta}_k(t)\right) \\
\phi_2 &= f_i\left(\tilde{i}_k(t) + \frac{\phi_1}{2},\ \tilde{\Theta}_k(t) + \frac{\phi_1}{2}\right) \\
\phi_3 &= f_i\left(\tilde{i}_k(t) + \frac{\phi_2}{2},\ \tilde{\Theta}_k(t) + \frac{\phi_2}{2}\right) \\
\phi_4 &= f_i\left(\tilde{i}_k(t) + \phi_3,\ \tilde{\Theta}_k(t) + \phi_3\right)
\end{aligned}$$

with which we can obtain

$$\begin{aligned}
\tilde{r}_k(t) &= \gamma \tilde{i}_k(t) + \tilde{r}_k(t-1) \\
\tilde{s}_k(t) &= 1 - \tilde{i}_k(t) - \tilde{r}_k(t)
\end{aligned}$$

for each step of t.

be the average degree of each group, then a feasible sampling strategy would to sample $\tilde{N}^{\tau*}_{k_h}(t)$, $\tilde{N}^{\tau*}_{k_m}(t)$, and $\tilde{N}^{\tau*}_{k_l}(t)$ agents from each strata at each time t.

3.5 Simulation

In this section, we show how the sampling strategies perform through simulations of sampling agents in an epidemic over generated random networks. We present the results focused on the Infectious (I) state as it is common to be of more interest than the other two states.[9]

3.5.1 Network and Epidemics

We generate a set of 100 random networks with size $\mathcal{N} = 10,000$ each using the configuration model proposed by Barabasi (2006). This method requires a predefined degree sequence, which we generate via a log-normal distribution. The average degree mean and standard deviation of the simulated networks are approximately 20 and 25 respectively. Figure shows the average degree distribution of the simulated networks.

Next, we simulate 10,000 epidemics, where each is constructed as the following:

(1) At $t = 0$, three agents are infected randomly. This determines the initial SIR vectors $x^S(0), x^I(0), x^R(0)$ of size $\mathcal{N}$ as $x^S_j(1) = 1 - x^I_j(0) - x^R_j(0)$ and $x^R_j(0) = 0$ for all j.

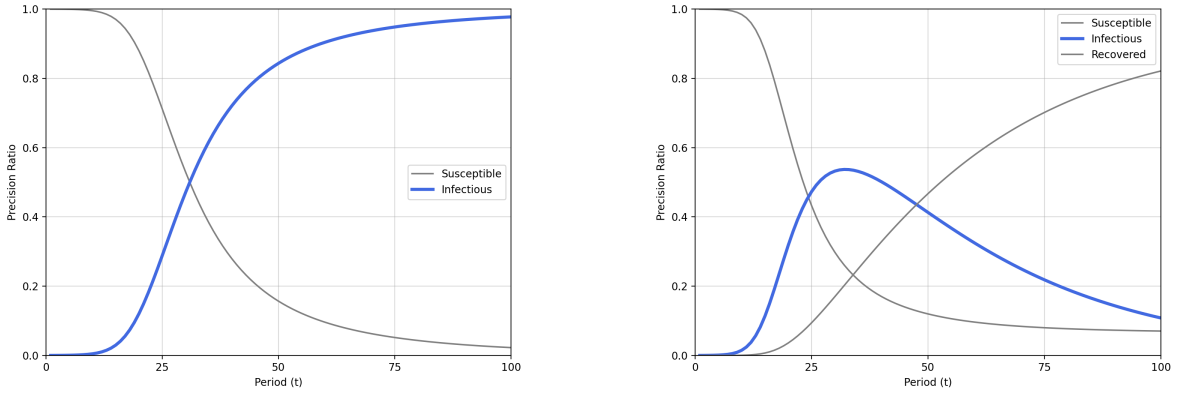
(2) For $t > 0$, agents in each state are redefined as, in chronological order,

- Susceptible (S): Susceptible agents at time t are those who were never infected in the previous periods, i.e. $x^S(t) = x^S(t-1) - x^I(t-1)$.
- Infectious (I): Each infected agent j (with $x^I_j(t) = 1$) infects their susceptible neighbor l (with $x^S_l(t) = 1$) with probability $\beta \in (0, 1)$ (*transmission rate*). If infected, $x^I_l(t+1) = 1$ and $x^S_l(t+1) = 0$ and vice versa if otherwise.

[9] Results focusing on Susceptible (S) and Recovered (R) states are in the Appendix.

a SI Model

b SIR Model



Note: The proportions of S,I, and R are the mean values of the 10,000 simulated epidemics over each 100 networks.

Figure 3.1: Epidemic Process of Each Model

- Recovered (R): Each infected agent j ($x_j^I(t) = 1$) recovers with probability γ (*recovery rate*). If recovered $x_j^I(t+1) = 0$ and $x_j^R(t+1) = 1$ and vice versa if otherwise.[10]

(3) The process advances up to until there are no infectious agents left or $t = 100$.

Figure 3.1 shows the average of simulated epidemics for each SI and SIR model. We chose $\beta = 0.01$ and $\gamma = 0.03$ for the SI Model and $\beta = 0.015$ and $\gamma = 0.03$ for the SIR model for the hyper parameters of the simulation.

3.5.2 Stratification and Sampling Schemes

For each simulated network, we divide the population into three groups – at the $2/3$ and $1/3$ percentile points of the degree distribution and name each group as *high*, *mid*, and *low*. Given

[10]For the SI Model, we simply discard the recovered (R) proportion of agents by setting $\gamma = 0$.

such, for each epidemic, we sample 1,000 agents (10% of the population) with three sampling schemes where the number of samples for each group $v \in \{k_{high}, k_{mid}, k_{low}\}$ in each period t for each epidemic state τ is:

[1] Optimal Stratified Sampling

$$N_v^{\tau *}(t) = \frac{q_v \sigma_{\epsilon,v}^{\tau}(t)}{\sum_w q_w \sigma_{\epsilon,w}^{\tau}(t)} N$$

[2] Feasible Stratified Sampling

$$\tilde{N}_v^{\tau}(t) = \frac{q_v \tilde{\sigma}_{\epsilon,v}^{\tau}(t)}{\sum_w q_w \tilde{\sigma}_{\epsilon,w}^{\tau}(t)} N$$

[3] Stratified Random Sampling

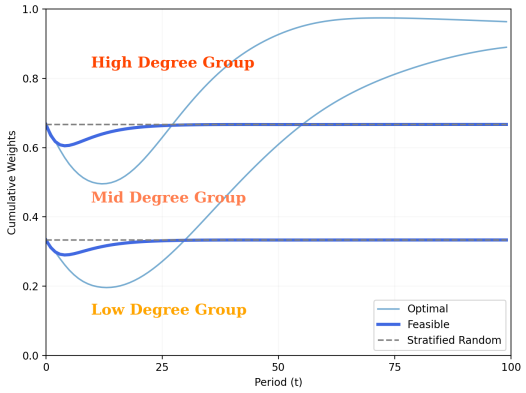
$$\hat{N}_v^{\tau}(t) = \frac{q_v}{\sum_w q_w} N$$

Note that the feasible scheme could be predefined since $\tilde{\sigma}_{\epsilon,v}^{\tau}(t)$ is obtained by a set of differential equations (3.3) and (3.4) through the Runge-Kutta method with known β and γ. For the optimal stratified sampling, we initially run 1,000 epidemics to find a proxy value (simulation estimated value) for the true within-strata variance, $\sigma_{\epsilon,v}^{\tau}(t)$, for each period. For the stratified random sampling, the weights are simply the proportion of the size of the groups (we set value 1 for the standard deviations). Given each sampling scheme, we collect a total sum of 1,000 samples and calculate the weighted sample mean such as given in equation (3.1).

3.5.3 Performance

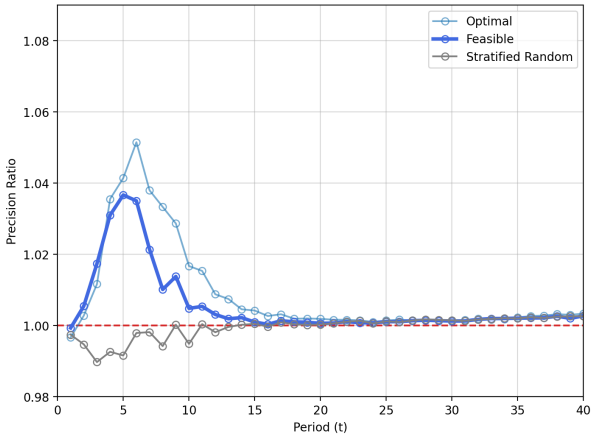
Given the samples for each simulation and network, we measure the performance of by comparing each method to simple random sampling of the same iteration. Specifically, for each sampling method, we obtain the simulation precision – inverse of simulation variance give then true mean from the total sample for each iteration – and construct a ratio based on the precision value from

a Std. Dev. Proportion of I by Group



Note: Each line represents the cumulative values of the standard deviation proportion by group. For example, at $t = 5$, the standard deviation in the high degree group (40% proportion) is roughly 25% greater than those from the mid degree (30% proportion).

b Sampling Performance – I ($\tau = I$)



Note: Each line represents the precision ratio of each sampling method to random sampling. If the value exceeds 1, then the method is more efficient than random sampling by the percentage amount it exceeds, e.g. at $t = 5$, feasible sampling is more efficient than random sampling by roughly 3.5%. Thus, both optimal and feasible sampling is more efficient than simple random sampling.

Figure 3.2: SI Model Results

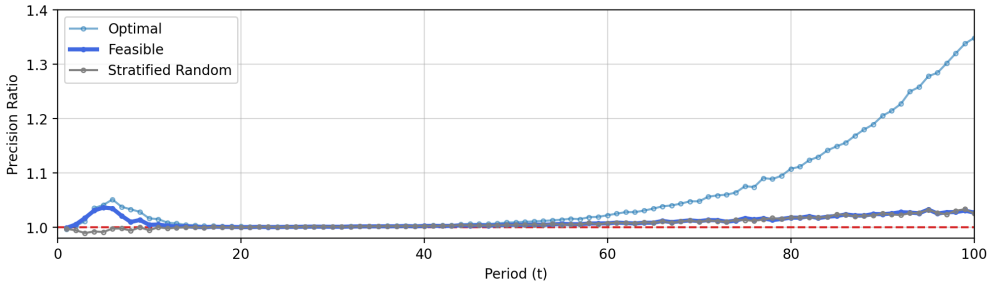
random sampling. Thus, if the precision ratio of each sampling method exceeds 1, then it is more efficient than random sampling by the magnitude of the difference.

SI Model

Figure 3.2a and the depicts the cumulative proportion of standard deviations for each sampling method of the SI model for $\tau = I$.[11] We can see that in both the optimal and feasible methods for

[11] We omit the SI model for $\tau = S$ since the results are the same as $\tau = I$ due to their binary nature; $Var(x^S(t)) = Var(x^I(t))$. The standard deviation proportion for $\tau = I$ is simply

$$\frac{\sigma^I_{\epsilon,v}(t)}{\sum_w \sigma^I_{\epsilon,w}(t)}$$



Note: Each line represents the precision ratio of each sampling method to random sampling. If the value exceeds 1, then the method is more efficient than random sampling by the percentage amount it exceeds.

Figure 3.3: Sampling Performance for All Periods – I ($\tau = I$)

all cases, the high degree group has a higher standard deviation value in the early periods, which shows that well connected agents are more prone to changes (i.e. $S \to I$) in the earlier stage. We also see that the standard deviation for the optimal case – or true $\sigma^{\tau}_{\epsilon,v}(t)$ – tends to be greater in the low degree group towards the end of the epidemic.

Another point we observe is that when the proportion of each state "cross" – i.e. when the proportion of I becomes larger than that of S for the SI model – we see that the optimal method lines cross those of the stratified random sampling as well.[12] That is, the variance in each group become the same at the "cross". From Figure 3.1a, we can see that this happens near period 30 ($t = 30$) for the SI model. This provides some guidance on how one can setup a sampling strategy when very minimal information is given; e.g. one can simply have a high degree biased sampling scheme before the "cross" then switch to a low degree biased scheme.

Figure 3.2b shows the performance of the sampling methods in the earlier stages ($t \leq 40$). We first focus on the early stage as it is the region before the "cross". As shown in Figure 3.2b, optimal sampling dominates all other methods almost everywhere though feasible sampling follows

for all $v \in \{k_{high}, k_{mid}, k_{low}\}$

[12]This "cross" does not necessarily exist as it could be a point where the decreasing rate of proportion of susceptibles (S) start to decrease (i.e. switch of sign of the second derivative). We continue to use the term "cross" as it is easier to describe with the given figures.

closely. Both methods peak around period $t = 5$ and $t = 6$ and slowly starts to decline. We can see that the optimal sampling method is consistently dominant even after period $t = 20$, however all three methods converge around the "cross" point, i.e. period $t = 30$. Note that stratified random sampling performs poorly compared to other methods in up to $t = 20$. This is due to sampling unnecessary agents from the low degree group which one can assume to have agents with a very low chance of being infected in the early stages.

Figure 3.3 shows the performance of the sampling methods for all stages ($t = 1, ..., 100$). As we focus on the phenomena after the "cross", we see that there is a gradual performance gain for all methods after period $t = 30$ and a further exponential gain for the optimal sampling. This could be expected from what we observe in Figure 3.2a as there is very little transition ($S \rightarrow I$) in the *high* degree group after period $t = 50$, meaning that almost all agents in that group are already infected. However, since random sampling continues to sample those agents, thus becoming a less efficient method. In a similar sense, feasible and stratified random sampling is slightly better than simple random sampling, though not at the level of optimal sampling, as it over samples unnecessary agents.

SIR Model

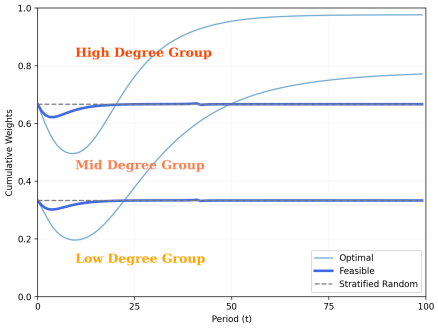
Contrasting to the design of the SI Model where we only considered one variance base (the other follows due to the binary nature), we have three separate variances to consider for the SIR Model. In other words, the variance of S for each group is $Var(x^S(t) * I_{k_h})$, $Var(x^S(t) * I_{k_m})$, and $Var(x^S(t) * I_{k_l})$ where I_k is an indicator vector where it takes value 1 in the indicies for the corresponding group and 0 otherwise. We can obtain the same sets for the case of I and R and thus have 9 combinations of cases total.[13] In Figure 3.4, we report the results for the corresponding variance bases.[14]

The first column in Figure 3.4 shows the cumulative proportion of standard deviations for each

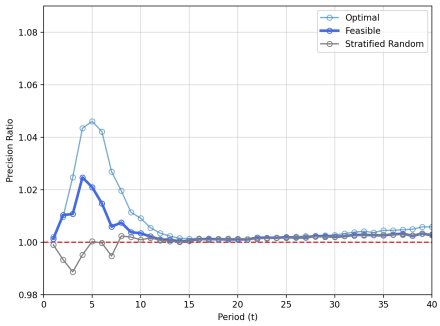
[13]We show the results of all 9 cases in Figures C.1, C.2, C.3, and C.4 in Appendix C.2.

[14]One caveat in applying the method for the SIR model (or a model with more than three states) is that the researcher must select a state of interest to choose the corresponding variance based bias scheme.

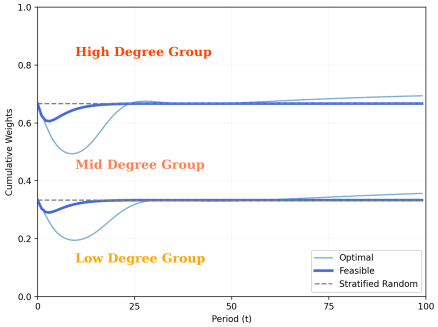
a Std. Dev. Proportion of S by Group



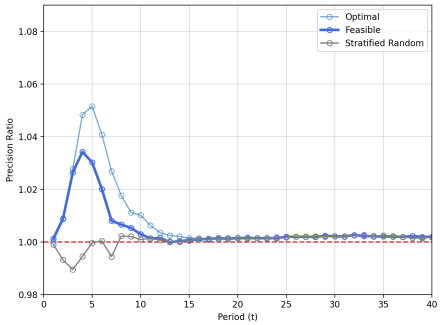
b Sampling Performance – S ($\tau = S$)



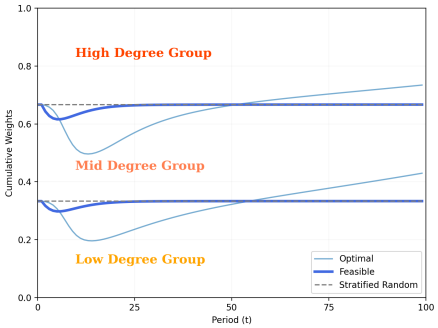
c Std. Dev. Proportion of I by Group



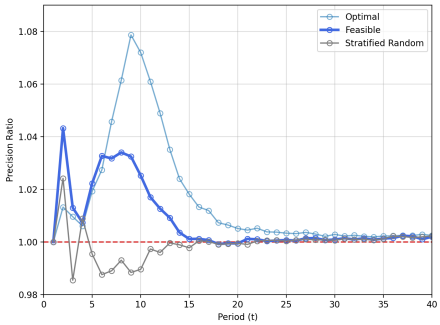
d Sampling Performance – I ($\tau = I$)



e Std. Dev. Proportion of R by Group



f Sampling Performance – R ($\tau = R$)



Note: Each line represents the cumulative values of the standard deviation proportion by group.

Note: Each line represents the precision ratio of each sampling method to random sampling. If the value exceeds 1, then the method is more efficient than random sampling by the percentage amount it exceeds.

Figure 3.4: SIR Model Results in Early Periods

sampling method for all three states. The proportions for S resembles that of the case in the SI model, while the other two show some similar patterns in the earlier stages. This suggests that early stage transitions are more concentrated among the highly connected agents. We also see the pattern in regards to the "cross" points of that of the SIR model shown in Figure 3.1b, where the lines for S and I cross around period $t = 25$, lines for S and R cross around period $t = 30$, and lines for I and R cross around period $t = 50$.

The second columns in Figure 3.4 shows the performance for each sampling method for each state with their corresponding variance base in the early periods $(t \leq 40)$. Similar to the results for the SI model, the optimal method dominates almost everywhere and the feasible method is more efficient than stratified random sampling.[15] One contrasting result, as shown in Figure 3.5 is that even in the latter periods, the optimal sampling method does not gain more efficiency as it has in the SI model case, as there are more than two states each agent can belong to.[16]

3.6 Remarks

We find that the there exists a non-feasible optimal sampling method that dominates all other methods. We further suggest a feasible method that dominates stratified random sampling and is as efficient as optimal sampling in some cases. We find that these efficiency gains are mostly in the early stages when sampling agents from groups with a low chance of being infected is unnecessary as it does not provide useful information. We argue that these findings can shed light on implementing sampling strategies in a real world epidemic scenario.[17]

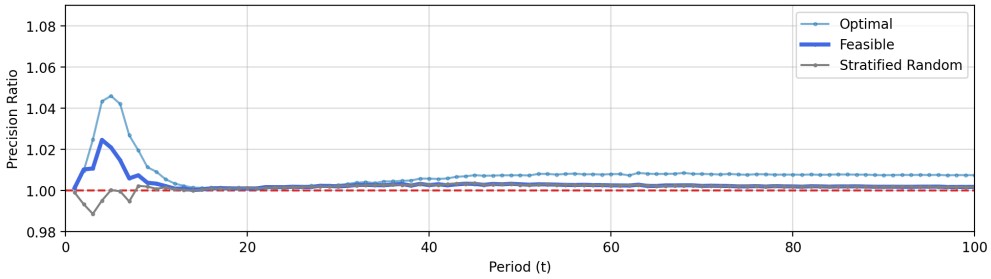
---

[15]The optimal method is less efficient than the other two in first couple periods for state R which is due to simulation noise as we only run 100 networks.

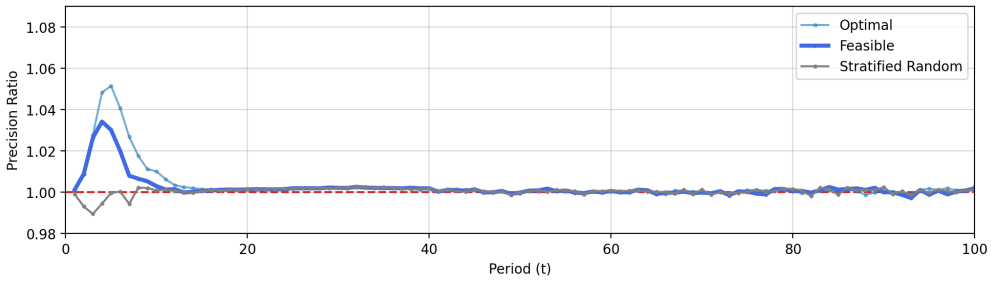
[16]We expect to see a higher gain for further periods: when the proportion of all of each states converge to it's limit.

[17]In a real world scenario, our method could be utilized with minimal resources. For example of a aerosol based disease epidemic in a region, we can divide the population to two groups depending on the average hours each individual spends in a populated area (i.e. public transportation, mall, library etc.) per a certain period of time (i.e. daily, weekly, etc.). Then, we can over-sample the individuals in the *high* hours group in the early periods up to the "cross" or where the decline of proportion of susceptible individuals start to slow down.

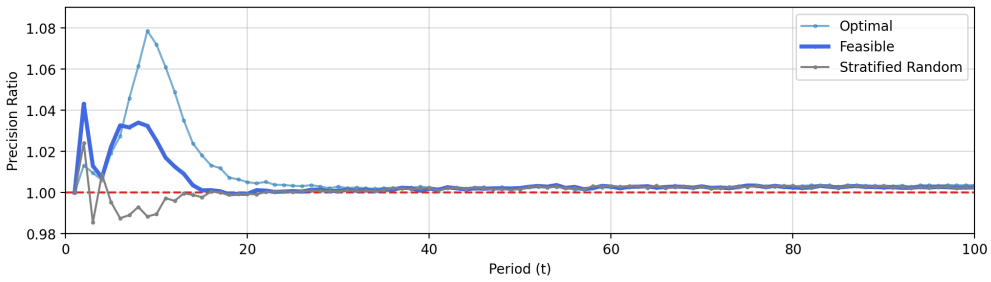
a Sampling Performance – S ($\tau = S$)



b Sampling Performance – I ($\tau = I$)



c Sampling Performance – R ($\tau = R$)



Note: Each line represents the precision ratio of each sampling method to random sampling.

Figure 3.5: SIR Model Results for all Periods

BIBLIOGRAPHY

Angrist, Joshua. 2014. "The Perils of Peer Effects." *Labour Economics*, 30: 98–108.

Artiles, Mayra, and Holly Matusovich. 2022. "Doctoral Advisor Selection in Chemical Engineering: Evaluating Two Programs through Principal-Agent Theory." *Studies in Engineering Education*, 2(2): 120.

Azoulay, Pierre, Joshua S. Graff Zivin, and Jialan Wang. 2010. "Superstar Extinction." *The Quarterly Journal of Economics*, 125(2): 549–589.

Bailey, Michael, Drew Johnston, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. 2022. "Peer effects in product adoption." *American Economic Journal: Applied Economics*, 14(3): 488–526.

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

Barabási, Albert-László. 2016. *Network Science.* Cambridge University Press.

Barnes, Benita J., and Ann E. Austin. 2009. "The Role of Doctoral Advisors: A Look at Advising from the Advisor's Perspective." *Innovative Higher Education*, 33(5): 297–315.

Beaman, Lori. 2011. "Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettled in the U.S." *Review of Economic Studies*, 79: 128–161.

Bergstrom, Carl T., Jevin D. West, and Marc A. Wiseman. 2008. "The Eigenfactor™ Metrics: Figure 1." *The Journal of Neuroscience*, 28(45): 11433–11434.

Blackwell, Matthew, Nicole E. Pashley, and Dominic Valentino. 2023. "Batch Adaptive Designs to Improve Efficiency in Social Science Experiments." *Working Paper*.

Blume, Lawrence E., William A. Brock, Steven N. Durlauf, and Rajshri Jayaraman. 2015. "Linear Social Interactions Models." *Journal of Political Economy*, 123(2): 444–496.

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

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

Bramoullé, Yann, and Sebastian Maes. 2024. "Measurement error and peer effects in networks." Unpublished Working Paper.

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

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. 2020. "Peer effects in networks: A survey." *Annual Review of Economics*, 12: 603–629.

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.

Brogaard, Jonathan, Joseph Engelberg, and Christopher A. Parsons. 2014. "Networks and productivity: Causal evidence from editor rotations." *Journal of Financial Economics*, 111(1): 251–270.

Brown, Kristine M., and Ron A. Laschever. 2012. "When They're Sixty-Four: Peer Effects and the Timing of Retirement." *American Economic Journal: Applied Economics*, 4(3): 90–115.

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.

Cai, Yong. 2022. "Regression with Centrality Measures." Unpublished Working Paper.

Cai, Yong, and Ahnaf Rafi. 2024. "On the performance of the Neyman Allocation with small pilots." *Journal of Econometrics*, 242(1): 105793.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications.* New York:Cambridge University Press.

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. 2016. *Econometrics of Network Formation.* Oxford University Press.

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

Clark, Samuel J., and Abigail Norris Turner. 2021. "Monitoring epidemics: Lessons from measuring population prevalence of the coronavirus." *Proceedings of the National Academy of Sciences*, 118(9).

Colussi, Tommaso. 2018. "Social Ties in Academia: A Friend Is a Treasure." *The Review of Economics and Statistics*, 100(1): 45–50.

Conti, Gabriella, Andrea Galeotti, Gerrit Mueller, and Stephen Pudney. 2013. "Popularity." *Journal of Human Resources*, 48(4): 1072–1094.

Corno, Lucia. 2017. "Homelessness and crime: do your friends matter?" *The Economic Journal*, 127(602): 959–995.

Dai, Jessica, Paula Gradu, and Christopher Harshaw. 2023. "Clip-OGD: An Experimental Design for Adaptive Neyman Allocation in Sequential Experiments." *Working Paper*.

de Paula, Áureo. 2020. "Econometric models of network formation." *Annu. Rev. Econom.*, 12(1): 775–799.

de Paula, Áureo, Imran Rasul, and Pedro Souza. 2020. "Identifying Network Ties from Panel Data: Theory and an Application to Tax Competition." Unpublished Working Paper.

de Paula, Áureo. 2017. "Econometrics of Network Models." In *Advances in Economics and Econometrics*. , ed. Bo Honore, Ariel Pakes, Monika Piazzesi and Larry Samuelson, 268–323. Cambridge:Cambridge University Press.

Dericks, Gerard, Edmund Thompson, Margaret Roberts, and Florence Phua. 2019. "Determinants of PhD student satisfaction: the roles of supervisor, department, and peer qualities." *Assessment & Evaluation in Higher Education*, 44(7): 1053–1068.

Ductor, Lorenzo, Marcel Fafchamps, Sanjeev Goyal, and Marco J. van der Leij. 2014. "Social Networks and Research Output." *The Review of Economics and Statistics*, 96(5): 936–948.

Ductor, Lorenzo, Sanjeev Goyal, and Anja Prummer. 2021. "Gender and Collaboration." *The Review of Economics and Statistics*, 1–40.

Epple, Dennis, and Richard Romano. 2011. "Peer Effects in Education: A Survey of the Theory and Evidence." In *Handbook of Social Economics*. Vol. 1, 1053–1165.

Fafchamps, Marcel, Sanjeev Goyal, and Marco J. van der Leij. 2010. "Matching and Network Effects." *Journal of the European Economic Association*, 8(1): 203–231.

Fu, J Sophia, Zhenghui Sha, Yun Huang, Mingxian Wang, Yan Fu, and Wei Chen. 2017. "Two-stage modeling of customer choice preferences in engineering design using bipartite network analysis." Vol. 58127, V02AT03A039, American Society of Mechanical Engineers.

Garber, Steven, and Steven Klepper. 1980. "Extending the classical normal errors-in-variables model." *Econometrica*, 1541–1546.

García-Suaza, Andrés, Jesús Otero, and Rainer Winkelmann. 2020. "Predicting early career productivity of PhD economists: Does advisor-match matter?" *Scientometrics*, 122(1): 429–449.

Gaule, Patrick, and Mario Piacentini. 2018. "An advisor like me? Advisor gender and post-graduate careers in science." *Research Policy*, 47(4): 805–813.

Goyal, Sanjeev, Marco J. van der Leij, and José Luis Moraga-González. 2006. "Economics: An Emerging Small World." *Journal of Political Economy*, 114(2): 403–412.

Graham, Bryan S. 2015. "Methods of identification in social networks." *Annu. Rev. Econom.*, 7(1): 465–485.

Griffith, Alan. 2022. "Name your friends, but only five? the importance of censoring in peer effects estimates using social network data." *Journal of Labor Economics*, 40(4): 779–805.

Griffith, Alan, and Sida Peng. 2023. "Identification of Network Structure in the Presence of Latent, Unobserved Factors: A New Result using Turán's Theorem." Unpublished Working Paper.

Gupta, Harsh, and Mason A Porter. 2022. "Mixed logit models and network formation." *Journal of Complex Networks*, 10(6): cnac045.

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.

Hilmer, Christiana, and Michael Hilmer. 2007. "Women Helping Women, Men Helping Women? Same-Gender Mentoring, Initial Job Placements, and Early Career Publishing Success for Economics PhDs." *American Economic Review*, 97(2): 422–426.

Hilmer, Michael J., and Christiana E. Hilmer. 2009. "Fishes, Ponds, and Productivity: Student-Advisor Matching and Early Career Publishing Success for Economics PhDs." *Economic Inquiry*, 47(2): 290–303.

Ho, Cheuk Yin. 2016. "Better health with more friends: the role of social capital in producing health." *Health Economics*, 25(1): 91–100.

Hsieh, Chih-Sheng, Michael D Koenig, Xiaodong Liu, and Christian Zimmermann. 2022. "Collaboration in Bipartite Networks." National Taiwan University, Department of Economics Working Papers 2202.

Hsieh, Chih-Sheng, Stenley I. M. Ko, Jaromir Kovarik, and Trevon D. Logan. 2024. "Non-randomly sampled networks: biases and corrections." *Journal of Econometrics*. Forthcoming.

Lee, Lung-fei. 2007. "Identification and Estimation of Econometric Models with Group Interactions, Contextual Factors, and Fixed Effects." *Journal of Econometrics*, 140(2): 333–374.

Lewbel, Arthur, Xi Qu, and Xun Tang. 2022. "Estimating Social Networks with Missing Links." Unpublished Working Paper.

Lewbel, Arthur, Xi Qu, and Xun Tang. 2023*a*. "Ignoring Measurement Errors in Social Networks." Unpublished Working Paper.

Lewbel, Arthur, Xi Qu, and Xun Tang. 2023*b*. "Social Networks with Unobserved Links." *Journal of Political Economy*, 898–946.

Litalien, David, and Frédéric Guay. 2015. "Dropout intentions in PhD studies: A comprehensive model based on interpersonal relationships and motivational resources." *Contemporary Educational Psychology*, 41: 218–231.

Liu, Xiaodong. 2013. "Estimation of a local-aggregate network model with sampled networks." *Economics Letters*, 118(1): 243–246.

Lleras-Muney, Adriana, Matthew Miller, Shuyang Sheng, and Veronica T Sovero. 2023. "Party on: The labor market returns to social networks and socializing." NBER Working Paper 27337.

Lovitts, Barbara E. 2001. *Leaving the ivory tower: The causes and consequences of departure from doctoral study.* Rowman & Littlefield.

Manresa, Elena. 2016. "Estimating the Structure of Social Interactions Using Panel Data." Unpublished Working Paper.

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

McPherson, Miller, Lynn Smith-Lovin, and James M Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology*, 27(1): 415–444.

Miguel, Edward, and Michael Kremer. 2004. "Worms: identifying impacts on education and health in the presence of treatment externalities." *Econometrica*, 72(1): 159–217.

Munshi, Kaivan. 2003. "Networks in the Modern Economy: Mexican Migrants in the U.S. Labor Market." *Quarterly Journal of Economics*, 118: 549–599.

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society*, 97(4): 558–625.

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.

Overgoor, Jan, Austin R. Benson, and Johan Ugander. 2020. "Choosing to Grow a Graph: Modeling Network Formation as Discrete Choice." arXiv:1811.05008 [physics].

Paré, Philip E., Carolyn L. Beck, and Tamer Başar. 2020. "Modeling, estimation, and analysis of epidemics over networks: An overview." *Annual Reviews in Control*, 50: 345–360.

Pezzoni, Michele, Jacques Mairesse, Paula Stephan, and Julia Lane. 2016. "Gender and the Publication Output of Graduate Students: A Case Study." *PLOS ONE*, 11(1): e0145146.

Plug, Erik, Bas van der Klaaw, and Lnnart Ziegler. 2018. "Do Parental Networks Pay Off? Linking Children's Labor-Market Outcomes to Their Parents' Friends." *Scandanavian Journal of Economics*, 120(1): 268–295.

Sacerdote, Bruce. 2011. "Peer Effects in Education: How Might They Work, How Big Are They, and How Much Do We Know Thus Far?" In *Hanbook of the Economics of Education*. Vol. 3, 249–277.

Sampson, Gabriel S, and Edward D Perry. 2019. "The role of peer effects in natural resource appropriation–the case of groundwater." *American Journal of Agricultural Economics*, 101(1): 154–171.

Sauermann, Henry, and Michael Roach. 2012. "Science PhD Career Preferences: Levels, Changes, and Advisor Encouragement." *PLoS ONE*, 7(5): e36307.

Sauer, Raymond D. 1988. "Estimates of the Returns to Quality and Coauthorship in Economic Academia." *Journal of Political Economy*, 96(4): 855–866.

Scharf, Kimberley, and Sarah Smith. 2016. "Relational altruism and giving in social groups." *Journal of Public Economics*, 141: 1–10.

Shi, Ying, and James Moody. 2017. "Most Likely to Succeed: Long-run Returns to Adolescent Popularity." *Social Currents*, 4(1): 13–33.

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

Tol, Richard S. J. 2021. "Rise of the Kniesians: The professor-student network of Nobel laureates in economics." *arXiv:2012.00103 [econ, q-fin]*. arXiv: 2012.00103.

Tripathi, Gautam. 1999. "A matrix extension of the Cauchy-Schwarz inequality." *Economics Letters*, 63(1): 1–3.

Wichmann, Bruno, Minjie Chen, and Wiktor Adamowicz. 2016. "Social networks and choice set formation in discrete choice models." *Econometrics*, 4(4): 42.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, Massachusetts:MIT Press.

Yeung, Fiona Chehong. 2019. "Statistical Revealed Preference Models for Bipartite Networks." University of California, Los Angeles Ph.D. Dissertation.

Zhao, Chun-Mei, Chris M. Golde, and Alexander C. McCormick. 2007. "More than a signature: how advisor choice and advisor behaviour affect doctoral student satisfaction." *Journal of Further and Higher Education*, 31(3): 263–281.

Zhao, Jinglong. 2023. "Adaptive Neyman Allocation." *SSRN Electronic Journal*.

Appendix A

APPENDIX FOR CHAPTER 1

A.1 Formal Definition of Graphs and Formation Processes

A.1.1 Academic Social Network

Through out the paper, we consider two graphs. The first graph is an author-to-author, time varying unipartite multilayer undirected graph $\mathcal{G}(\mathcal{V}_t, \mathcal{E}_t)$ with an discrete time index t. This graph consists one node set $\mathcal{V}_t$ of size n_t, where each node in this set is considered an author. This set expands by time, in which our setting, could be seen as new Ph.D. graduates joining the community of economics authors. These authors can either be in coauthorship relation, advisors-advisees relation, or both. This is represented in three different types of edge sets: the coauthorship network edge set $\mathcal{E}_t^C$, genealogy network – advisor-advisee relationship – edge set $\mathcal{E}_t^G$, and the academic social network edge set $\mathcal{E}_t^A$. The weights on each element in the edge set, (i, j), respectively, are defined as

$$w_{ij,t}^C = \sum_s \left(\frac{a_{ij,t}^C}{t - T_{ij,s}^P + 1} \right)^{\frac{1}{c}} \qquad \text{for } t \geq T_{ij,s}^P$$

$$w_{ij,t}^G = \left(\frac{a_{ij,t}^G}{t - T_{ij}^G + 1} \right)^{\frac{1}{c}} \qquad \text{for } t \geq T_{ij}^G$$

and

$$w_{ij,t}^A = w_{ij,t}^C + w_{ij,t}^G$$

where for authors i and j, $T_{ij,s}$ denotes the time of encounter of authors for project s, $T_{ij,g}$ denotes the time of encounter of authors as advisor-advisee relationship. [1]. $a^C_{ij,t}$, $a^G_{ij,t}$, and $a^A_{ij,t}$ are elements for corresponding $n_t \times n_t$ adjacency matrices $A^C_{\mathcal{G}_t}$, $A^G_{\mathcal{G}_t}$, and $A^A_{\mathcal{G}_t}$, respectfully. Thus, for authors i and j, if $a^C_{ij,t} = 1$, they had coauthored a project at time t; if $a^G_{ij,t} = 1$, they formed an advisor-advisee relationship at time t; and if $a^A_{ij,t} = 1$, either or both events happened.

We choose this weighting scheme in the view of considering each publication or the advisor-advisee relationship as a *academic social encounter*. It is natural to think that each connection to be less stronger as time goes by, even for advisee-advisor relationships, unless they frequently cooperate for projects. This weighting scheme allows us to i) construct the academic social network by simply adding the weighted adjacency matrix of the two edge sets as they share the same measurement and ii) choose the rate of discount through parameter c, which allows us the flexibility to simply calculate the strength centrality of each author node with a large c if needed.[2] The first property is illustrated in Figure 1.5 where we can see that the projection of the two first layers of the network to the final layer is the academic social network of all authors in $\mathcal{V}_t$.

The second graph is an author-to-project, time varying bipartite undirected graph $\mathcal{H}(\mathcal{V}_t, \mathcal{P}_t, \mathcal{E}^P_t)$ where P_t is the set of projects of size p_t that each author in V_t participates at time t. Let us denote the bi-adjacency matrix of the author-to-project network as $B_{\mathcal{H}_t}$ which has dimensions $n_t \times p_t$ and for each element $b_{is,t}$ in matrix $B_{\mathcal{H}_t}$ takes value 1 if author i participates in project s and 0 otherwise. Then, since

$$a^C_{ij,t} = \mathbf{1}\left\{\left[B_{\mathcal{H}_t} \times B^{\mathbf{T}}_{\mathcal{H}_t} > 0\right]_{i,j}\right\}$$

assigning a weighting scheme for each edge, (i, s) in edge set $\mathcal{E}^P_t$ as

$$w^P_{is,t} = \left(\frac{b_{is,t}}{t - T_{is} + 1}\right)^{\frac{1}{2c}} \qquad \text{for } t > T_{is}$$

allows us to construct the weighted adjacency matrix of the coauthorship network, $[w^C_{ij,t}]_{n_t \times n_t}$,

[1] We do not observe these actual time so we use the year of publication or presentation for the projects and year of graduation for the empirical work

[2] We set $c = 1$ for simplicity through out the paper.

by taking the outer product of the weighted bi-adjacency matrix of the author-to-project network, $[w^P_{is,t}]_{n_t \times p_t}$. Thus, identifying the bi-adjacency matrix $B_{\mathcal{H}_t}$ and the adjacency matrix $A^G_{\mathcal{G}_t}$ with their corresponding weights allows us to fully characterize both graphs.

A.1.2 Genealogy Network Formation

At time $T_{ij,g}$, Each advisee author i selects an advisor author j from a potential pool of advisor authors $\mathcal{V}_i^{G_a}$ conditional on the pairwise characteristics as well as the advisors' characteristics and advisees' fixed effects. Our assumption is based on the nature of the genealogy tree where there are multiple advisee authors connected to one advisor author and that it is more common, at least in the field of economics, for a student to propose to a professor of their choice after observing their characteristics.[3] We also consider the fact that each advisee student has a limited number of advisors to choose from, constrained by both time and place.

We translate this process to an econometric model by employing an asymmetric conditional logit model which likelihood takes form of

$$P\left(advisee_i = j | j \in \mathcal{V}_i^{G_a}\right) = \frac{\exp(\alpha d_{j,t} + z'_{ij,t}\delta)}{\sum\limits_{k \in \mathcal{V}_i^{G_a}} \exp(\alpha d_{j,t} + z'_{ij,t}\delta)}$$

where the $advisee_i$ is the decision output by advisee-author i. Thus the left hand side denote the probability of advisee author i choosing advisor author j from advisee author i's pool of possible advisors $\mathcal{V}_i^{G_a}$ and establishing an advisor-advisee relationship, i.e. forming an edge in the edge set $\mathcal{E}_t^G$. Note that the advisee chooses only one advisor once, so there the decision is not time varying but the set of pool is, which depends on when the advisee makes the decision.

On the right hand side, the model includes two types of covariates. The first type is the advisors' individual characteristics that are time varying. Namely, denoted as $d_{j,t}$, we use the number of past students of advisor author j at the time of when advisee author i observes. This is an augmented form of a preferential attachment with fitness process as described in Overgoor, Benson

[3] We assume that the advisor author – student – has enough information that they know whether their proposal will be rejected or accepted.

and Ugander (2020) where the degree of the selected node enters the equation as $\alpha\log(d_{j,t})$.[4] Our variation can be seen as the degree minus one (the advisor's advisor), which we leave out since not all of the advisors' advisors are observed, thus could not be calculated. We also neglect the log transformation since there are cases where advisor-authors have no past students. Figure 1.2 in the main text illustrates the out degree of the genealogy network where the downward sloping linear trend supports the usage of this approach.[5]

The second type are advisor-advisee pairwise fixed effects. Note that the conditional logit model requires covariates to be varying in j so we cannot directly use fixed effects of advisees that only vary in i. However, we can measure the interaction term of a pairwise fixed effect by splitting the datasets into blocks at the cost of efficiency. This is equivalent with multiplying each pairwise fixed effect variable with corresponding dummy variables thus creating a matrix of the dimension of the number of categories. For example, for a given pairwise binary covariate such as gender, we can construct two pairwise covariates of same gender (1 if advisor-advisee have the same gender, 0 if not), one for the male advisees and other for the female advisees and thus measure the difference of preferences by gender.

A.1.3 Coauthorship Network Formation

As shown in Section A.1.1, the coauthorship network is fully identified by the weighted bi-adjacency matrix of the author-to-project network, $B_{\mathcal{H}_t}$. In order to model the growth process of this bipartite network, we take a similar approach as the genealogy network case by assuming the advisee authors choose a project from a pool of possible projects they can participated in, conditional on the characteristics of the other participating coauthors. A key difference is that we allow the advisee authors to choose multiple projects regardless of chronological order. We simply assume each advisee-author chooses projects conditional on their genealogy network and pairwise

[4] A preferential attachment model in this setup would have a probability of $P(h_i = j | j \in \mathcal{V}_i^{G_a}) = \frac{d_{j,t}^{\alpha}}{\sum_{k \in \mathcal{V}_i^{G_a}} d_{k,t}^{\alpha}}$.

[5] The degree distribution of a network built from a preferential attachment process will have a pareto distriubtion, thus having a down ward sloping linear trend of the log-log plot.

fixed effects from that of with the corresponding coauthors. This assumption relies on the fact that we are only focusing on the first couple projects of the advisee-author, more so on those that they have participated in during graduate school. Thus, the selection process would more likely be effected by the initial academic social encounters; their advisors and advisors' coauthors. This situation also makes it difficult to correctly distinguish the choronological order of projects, hence the myopic setup.

In order to allow the conditional logit model to accompany cases with multiple choices from asymmetric multiple categories, we assume the choices are independent among individuals. Then, the likelihood of the model takes form of

$$P\left(advisee_i = s_n | s_n \in \mathcal{P}_i, \mathcal{G}\right) = \prod_{s_n}^{r_i} \frac{\exp(q'_{is_n,t}\theta)}{\sum\limits_{k \in \mathcal{P}_{i,t}} \exp(q'_{ik,t}\theta)}$$

where $advisee_i$ is the choice vector, i.e. the decision made by advisee-author i who can choose projects s_n, for $n \in \{1, ..., r_i\}$ where r_i is the number of projects advisee i participates in, from their pool of projects $\mathcal{P}_{i,t}$. Thus the left hand side is the probability of advisee-author i joining r_i number of projects, represented as s, conditional on the pairwise characteristics between the corresponding coauthors represented by $q_{i_n s,t}$. Note that, denoted by r_i, the number of projects each advisee-author i joins could be regarded as the capacity (as referred to in the main text) or ability of the advisee-author and we assume that this not conditional on any information and thus fixed.

A.2 Additional Tables

Table A.1: Proportion of Generated Statistics Greater than True Values

	1st order degree centrality		2nd order degree centrality	
	mean	median	mean	median
Case 1	0.9734	0.7893	0.8111	0.3971
Case 2	0.9709	0.7893	0.8015	0.3656
Case 3	0.9709	0.8111	0.7990	0.3680

Note: These are results show the proportion of where the median or mean of the draws of the networks statistics for each case, exceed the true network statistic value for each advisee sample.

Table A.2: Estimated Mean Regressions Results for the Log-Linear Production Function

	Dependent variable: $\log \bar{y}_i$				
	(1)	(2)	(3)	(4)	(5)
Advisee Male	0.221	0.277	0.165	0.227	0.195
	(0.159)	(0.534)	(0.508)	(0.511)	(0.155)
Advisor Male		0.197	0.126	0.162	
		(0.506)	(0.482)	(0.488)	
Both Male		-0.061	0.027	-0.036	
		(0.568)	(0.541)	(0.542)	
1st order Degree Centrality			2.101***	2.587***	2.611***
			(0.354)	(0.600)	(0.611)
2nd order Degree Centrality				-0.394	-0.413
				(0.467)	(0.474)
constant	1.084*	0.894	0.593	0.524	0.680
	(0.562)	(0.734)	(0.704)	(0.705)	(0.540)
Eastern Africa	-0.167	-0.164	-0.202	-0.244	-0.249
	(0.697)	(0.691)	(0.702)	(0.708)	(0.715)
Eastern Asia	-1.474***	-1.428***	-1.494***	-1.485***	-1.526***
	(0.497)	(0.511)	(0.501)	(0.502)	(0.490)
Eastern Europe	-0.684	-0.669	-0.752	-0.740	-0.755
	(0.497)	(0.501)	(0.486)	(0.486)	(0.483)
Northern Africa	-0.986*	-0.927*	-1.302***	-1.263**	-1.315***
	(0.512)	(0.528)	(0.489)	(0.495)	(0.486)
Northern America	-0.515	-0.493	-0.502	-0.506	-0.527
	(0.474)	(0.479)	(0.470)	(0.471)	(0.467)
Northern Europe	-0.255	-0.244	-0.220	-0.222	-0.232
	(0.556)	(0.562)	(0.566)	(0.567)	(0.562)
South America	-0.663	-0.633	-0.702	-0.697	-0.723
	(0.555)	(0.564)	(0.554)	(0.556)	(0.547)
South-eastern Asia	-0.703	-0.694	-0.553	-0.529	-0.536
	(0.810)	(0.817)	(0.759)	(0.753)	(0.747)
Southern Asia	-0.370	-0.357	-0.432	-0.435	-0.447
	(0.511)	(0.515)	(0.503)	(0.503)	(0.501)
Southern Europe	-0.698	-0.680	-0.720	-0.720	-0.736
	(0.489)	(0.491)	(0.482)	(0.483)	(0.482)
Western Africa	-0.668	-0.604	-0.925	-0.896	-0.953
	(0.939)	(0.967)	(0.640)	(0.644)	(0.619)
Western Asia	-0.490	-0.472	-0.423	-0.417	-0.432
	(0.489)	(0.496)	(0.486)	(0.487)	(0.480)
Western Europe	-0.553	-0.527	-0.534	-0.517	-0.540
	(0.471)	(0.478)	(0.468)	(0.470)	(0.464)
Observations	431	431	431	431	431
R^2	0.294	0.295	0.332	0.333	0.333
Adjusted R^2	0.184	0.181	0.221	0.221	0.224

*p<0.1; **p<0.05; ***p<0.01

Note: Extended table of Table1.5. Institution fixed effects are omitted. Standard errors in parentheses are heteroscedasticity robust standard errors.

Table A.3: Estimated Mean Regressions Results for the Linear Production Function

	Dependent variable: $\bar{y}_i$				
	(1)	(2)	(3)	(4)	(5)
Advisee Male	0.834**	0.946	0.491	0.877	0.742*
	(0.402)	(0.945)	(0.837)	(0.847)	(0.383)
Advisor Male		0.409	0.123	0.339	
		(0.913)	(0.839)	(0.844)	
Both Male		-0.123	0.236	-0.151	
		(1.106)	(0.999)	(0.987)	
1st order Degree Centrality			8.539***	11.531***	11.561***
			(2.076)	(2.728)	(2.750)
2nd order Degree Centrality				-2.422**	-2.443**
				(0.959)	(0.991)
constant	1.955	1.559	0.339	-0.086	0.235
	(1.439)	(1.552)	(1.499)	(1.557)	(1.478)
Eastern Africa	1.402	1.408	1.256	0.997	0.993
	(2.263)	(2.248)	(2.292)	(2.330)	(2.344)
Eastern Asia	-1.122	-1.026	-1.294	-1.238	-1.304
	(1.061)	(1.090)	(0.995)	(0.995)	(0.972)
Eastern Europe	-0.158	-0.125	-0.463	-0.392	-0.411
	(1.486)	(1.512)	(1.304)	(1.298)	(1.280)
Northern Africa	-1.068	-0.944	-2.469**	-2.230**	-2.312**
	(1.270)	(1.304)	(1.072)	(1.082)	(1.061)
Northern America	-0.010	0.036	0.002	-0.026	-0.057
	(1.108)	(1.127)	(1.053)	(1.044)	(1.030)
Northern Europe	0.388	0.413	0.508	0.498	0.482
	(1.355)	(1.365)	(1.362)	(1.358)	(1.351)
South America	-0.635	-0.572	-0.852	-0.818	-0.862
	(1.312)	(1.323)	(1.280)	(1.278)	(1.269)
South-eastern Asia	-0.571	-0.553	0.018	0.170	0.162
	(2.109)	(2.130)	(1.965)	(1.952)	(1.936)
Southern Asia	0.198	0.224	-0.081	-0.096	-0.111
	(1.377)	(1.390)	(1.303)	(1.303)	(1.295)
Southern Europe	0.082	0.120	-0.044	-0.040	-0.064
	(1.186)	(1.208)	(1.106)	(1.100)	(1.086)
Western Africa	4.929	5.063	3.758	3.936	3.848
	(5.846)	(5.848)	(4.097)	(4.116)	(4.113)
Western Asia	-0.095	-0.057	0.142	0.182	0.157
	(1.114)	(1.126)	(1.056)	(1.053)	(1.044)
Western Europe	0.238	0.293	0.264	0.368	0.333
	(1.131)	(1.155)	(1.073)	(1.074)	(1.058)
Observations	431	431	431	431	431
R^2	0.294	0.295	0.332	0.333	0.333
Adjusted R^2	0.184	0.181	0.221	0.221	0.224

*p<0.1; **p<0.05; ***p<0.01

Note: Estimated results of the linear production function (non-log-linear). Standard errors in parentheses are heteroscedasticity robust standard errors. Compared to the results of the log-linear model, we can see that the coefficient for the gender dummy variable is statistically significant, but at a 10% level for the model with the best goodness-of-fit. Institution fixed effect estimates are omitted.

Table A.4: Estimated Quantile Regressions Results for the Log-Linear Production Function

	Dependent variable: $\log \bar{y}_i$								
τ	$q_{0.1}$	$q_{0.2}$	$q_{0.3}$	$q_{0.4}$	$q_{0.5}$	$q_{0.6}$	$q_{0.7}$	$q_{0.8}$	$q_{0.9}$
Advisee Male	0.323*	0.239	0.111	0.151	0.139	0.130	0.175	0.137	0.031
	(0.180)	(0.161)	(0.140)	(0.124)	(0.125)	(0.121)	(0.121)	(0.126)	(0.127)
1st order Degree	3.519***	2.636***	2.135***	1.784***	1.569***	1.988***	1.427***	1.813***	3.281***
	(1.101)	(0.787)	(0.667)	(0.583)	(0.542)	(0.487)	(0.460)	(0.438)	(0.476)
2nd order Degree	-0.566	0.018	-0.098	-0.093	0.003	-0.158	0.080	-0.060	-0.658***
	(0.689)	(0.485)	(0.443)	(0.353)	(0.328)	(0.295)	(0.282)	(0.246)	(0.247)
Constant	0.044	-0.572	-0.281	-0.136	0.877	1.591***	1.435***	1.143*	0.504
	(1.057)	(0.995)	(0.749)	(0.621)	(0.606)	(0.591)	(0.543)	(0.587)	(0.763)
Observations	431	431	431	431	431	431	431	431	431
Pseudo R-squared	0.3841	0.2804	0.2264	0.2002	0.1968	0.1895	0.1919	0.2013	0.2466

*p<0.1; **p<0.05; ***p<0.01

Note: Estimate results of quantile regerssions. Standard errors in parentheses are heteroscedasticity robust standard errors. Institution and region of origin fixed effects are omitted.

Appendix B

APPENDIX FOR CHAPTER 2

B.1 Proofs

Lemma 1

For Parts [1] and [3], we derive an expression as a function of $\mathbf{W}$, then substitute the definition of $\mathbf{W}$ at the end.

As a preliminary matter, define $W_{ij}(L_{ij}, d_i^{true})$, noting that this is a deterministic function. Further, since $W_{ij}(0, d_i^{true}) = 0$ for any d_i^{true},

$$\mathbb{E}[W_{ij}|d_i^{true}] = Pr(L_{ij} = 1|d_i^{true})W_{ij}(1, d_i^{true}) \tag{B.1}$$

Since $L_{ij} \in \{0, 1\}$ and $W_{ij}(0, d_i^{true}) = 0$,

$$\mathbb{E}[W_{ij}\mathbf{x}_{(i)}(\mathbf{W}^{m-1}\mathbf{x})'_{(j)}|d_i^{true}] = Pr(L_{ij} = 1|d_i^{true})W_{ij}(1, d_i^{true})\mathbb{E}[\mathbf{x}_{(i)}(\mathbf{W}^{m-1}\mathbf{x})'_{(j)}|L_{ij} = 1, d_i^{true}] \tag{B.2}$$

Assumption 5, part [1] and Line (B.1) then imply that this can be simplified as

$$\mathbb{E}[W_{ij}\mathbf{x}_{(i)}(\mathbf{W}^{m-1}\mathbf{x})'_{(j)}|d_i^{true}] = \mathbb{E}[W_{ij}|d_i^{true}]\mathbf{A}_{m-1}(d_i^{true}) \tag{B.3}$$

And therefore,

$$\mathbb{E}[\sum_j W_{ij}\mathbf{x}_{(i)}(\mathbf{W}^{m-1}\mathbf{x})'_{(j)}|d_i^{true}] = \mathbb{E}[\sum_j W_{ij}|d_i^{true}]\mathbf{A}_{m-1}(d_i^{true}) \tag{B.4}$$

Next,

$$\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{W}^m\mathbf{x}] = \frac{1}{N}\sum_i \mathbb{E}[\mathbf{x}_{(i)}(\mathbf{W}^m\mathbf{x})'_{(i)}] = \mathbb{E}[\mathbf{x}_{(i)}(\mathbf{W}^m\mathbf{x})'_{(i)}] \tag{B.5}$$

$$= \mathbb{E}[\sum_j W_{ij}\mathbf{x}_{(i)}(\mathbf{W}^{m-1}\mathbf{x})'_{(j)}] \tag{B.6}$$

where the last substitution applies the fact that $(\mathbf{W}^m\mathbf{x})_{(i)} = \sum_j W_{ij}(\mathbf{W}^{m-1}\mathbf{x})_{(j)}$. By iterated expectations, this becomes

$$\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{W}^m\mathbf{x}] = \mathbb{E}[\mathbb{E}[\sum_j W_{ij}\mathbf{x}_{(i)}(\mathbf{W}^{m-1}\mathbf{x})'_{(j)}|d_i^{true}]] \tag{B.7}$$

Substitute (B.4) into (B.7) and simplify, which gives

$$\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{W}^m\mathbf{x}] = \mathbb{E}[\mathbb{E}[\sum_j W_{ij}|d_i^{true}]\mathbf{A}_{m-1}(d_i^{true})] \tag{B.8}$$

$$= \mathbb{E}[\sum_j W_{ij}\mathbf{A}_{m-1}(d_i^{true})] \tag{B.9}$$

When $\mathbf{W} = \mathbf{G}$, $\sum_j W_{ij} = 1$, which gives the result in part [1].

When $\mathbf{W} = \mathbf{L}$, $\sum_j W_{ij} = d_i^{true}$, which gives the result in part [3].

For Parts [2] and [4], we derive an expression as a function of $\mathbf{W}^*$, then substitute the definition of $\mathbf{W}^*$ at the end.

As a preliminary matter, note that W_{ij}^* is a deterministic function of M_{ij} and d_i^{obs}. Define $W_{ij}^*(M_{ij}, d_i^{obs})$. Since $W_{ij}^*(0, d_i^{obs}) = 0$ for any d_i^{obs},

$$\mathbb{E}[W_{ij}^*|d_i^{true}, d_i^{obs}] = Pr(M_{ij} = 1|d_i^{true}, d_i^{obs})W_{ij}^*(1, d_i^{obs}) \tag{B.10}$$

Since $M_{ij} \in \{0, 1\}$ and $W^*_{ij}(0, d_i^{obs}) = 0$,

$$\mathbb{E}[W^*_{ij}\mathbf{x}_{(i)}\mathbf{x})'_{(j)}|d_i^{true}, d_i^{obs}] = Pr(M_{ij} = 1|d_i^{true}, d_i^{obs})W^*_{ij}(1, d_i^{obs})\mathbb{E}[\mathbf{x}_{(i)}\mathbf{x}'_{(j)}|M_{ij} = 1, d_i^{true}, d_i^{obs}] \tag{B.11}$$

Assumption 5, part [1] and Line (B.10) then imply that this can be simplified as

$$\mathbb{E}[W^*_{ij}\mathbf{x}_{(i)}\mathbf{x}'_{(j)}|d_i^{true}, d_i^{obs}] = \mathbb{E}[W^*_{ij}|d_i^{true}, d_i^{obs}]\mathbf{A}_0(d_i^{true}) \tag{B.12}$$

In turn,

$$\mathbb{E}[W^*_{ij}\mathbf{x}_{(i)}\mathbf{x}'_{(j)}|d_i^{true}] = \left(\sum_d Pr(d_i^{obs} = d|d_i^{true})\mathbb{E}[W^*_{ij}|d_i^{true}, d_i^{obs}]\right)\mathbf{A}_0(d_i^{true}) \tag{B.13}$$

$$= \mathbb{E}[W^*_{ij}|d_i^{true}]\mathbf{A}_0(d_i^{true}) \tag{B.14}$$

Summing up for all j,

$$\mathbb{E}[\sum_j W^*_{ij}\mathbf{x}_{(i)}\mathbf{x}'_{(j)}|d_i^{true}] = \mathbb{E}[\sum_j W^*_{ij}|d_i^{true}]\mathbf{A}_0(d_i^{true}) \tag{B.15}$$

Next,

$$\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{W}^*\mathbf{x}] = \frac{1}{N}\sum_i \mathbb{E}[\mathbf{x}_{(i)}(\mathbf{W}^*\mathbf{x})'_{(i)}] = \mathbb{E}[\mathbf{x}_{(i)}(\mathbf{W}^*\mathbf{x})'_{(i)}] \tag{B.16}$$

$$= \mathbb{E}[\sum_j W^*_{ij}\mathbf{x}_{(i)}\mathbf{x}'_{(j)}] \tag{B.17}$$

where the last substitution applies the fact that $(\mathbf{W}^*\mathbf{x})_{(i)} = \sum_j W^*_{ij}\mathbf{x}_{(j)}$. By iterated expectations, this becomes

$$\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{W}^*\mathbf{x}] = \mathbb{E}[\mathbb{E}[\sum_j W^*_{ij}\mathbf{x}_{(i)}\mathbf{x}'_{(j)}|d_i^{true}]] \tag{B.18}$$

Substitute (B.15) into (B.18) and simplify, which gives

$$\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{W}^*\mathbf{x}] = \mathbb{E}[\mathbb{E}[\sum_j W_{ij}^* | d_i^{true}]\mathbf{A}_0(d_i^{true})] \tag{B.19}$$

$$= \mathbb{E}[\sum_j W_{ij}^* \mathbf{A}_0(d_i^{true})] \tag{B.20}$$

When $\mathbf{W} = \mathbf{G}$ (and thus $\mathbf{W}^* = \mathbf{H}$), $\sum_j W_{ij}^* = 1$. Substitution into (B.20) then gives the result in part [2].

When $\mathbf{W} = \mathbf{L}$ (and thus $\mathbf{W}^* = \mathbf{M}$), $\sum_j W_{ij}^* = d_i^{obs}$. Substitution into (B.20) then gives the result in part [4].

Lemma 2

As a preliminary matter, define $W_{ij}(L_{ij}, d_i^{true})$ and $W_{ij}^*(M_{ij}, d_i^{obs})$, noting that both are deterministic functions.

For Parts [1] and [4], we derive an expression as a function of $\mathbf{W}$, then substitute the definition of $\mathbf{W}$ at the end.

Since $W_{ij}(0, d_i^{true}) = 0$ for any d_i^{true},

$$\mathbb{E}[W_{ij}^2 | d_i^{true}] = Pr(L_{ij} = 1 | d_i^{true}) W_{ij}^2(1, d_i^{true}) \tag{B.21}$$

$L_{ij} \in \{0, 1\}$ then implies

$$\mathbb{E}[(W_{ij})^2 \mathbf{x}_{(j)}(\mathbf{W}^{m-1}\mathbf{x})'_{(j)} | d_i^{true}] = Pr(L_{ij} = 1 | d_i^{true})(W_{ij}(1, d_i^{true}))^2 \mathbb{E}[\mathbf{x}_{(j)}(\mathbf{W}^{m-1}\mathbf{x})'_{(j)} | L_{ij} = 1, d_i^{true}] \tag{B.22}$$

Assumption 5, part [2] and Line (B.21) then imply that this can be simplified as

$$\mathbb{E}[(W_{ij})^2 \mathbf{x}_{(j)}(\mathbf{W}^{m-1}\mathbf{x})'_{(j)} | d_i^{true}] = \mathbb{E}[W_{ij}^2 | d_i^{true}] \mathbf{B}_{m-1}(d_i^{true}) \tag{B.23}$$

And therefore,

$$\mathbb{E}[\sum_j (W_{ij})^2 \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(j)} | d_i^{true}] = \mathbb{E}[\sum_j W_{ij}^2 | d_i^{true}] \mathbf{B}_{m-1}(d_i^{true}) \tag{B.24}$$

Similarly, since $L_{ij}, L_{ik} \in \{0, 1\}$, whenever $k \neq j$,

$$\mathbb{E}[W_{ij} W_{ik} | d_i^{true}] = Pr(L_{ij} = 1, L_{ik} = 1 | d_i^{true}) W_{ij}(1, d_i^{true}) W_{ik}(1, d_i^{true}) \tag{B.25}$$

$L_{ij}, L_{ik} \in \{0, 1\}$ then implies

$$\begin{aligned}&\mathbb{E}[W_{ij} W_{ik} \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(k)} | d_i^{true}] \\ &\quad = Pr(L_{ij} = 1, L_{ik} = 1 | d_i^{true}) W_{ij}(1, d_i^{true}) W_{ik}(1, d_i^{true}) \mathbb{E}[\mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(k)} | L_{ij} = 1, L_{ik} = 1, d_i^{true}]\end{aligned} \tag{B.26}$$

Assumption 5, part [3] and Line (B.25) then imply that this can be simplified as

$$\mathbb{E}[W_{ij} W_{ik} \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(k)} | d_i^{true}] = \mathbb{E}[W_{ij} W_{ik} | d_i^{true}] \mathbf{C}_{m-1}(d_i^{true}) \tag{B.27}$$

and therefore

$$\mathbb{E}[\sum_j \sum_{k \neq j} W_{ij} W_{ik} \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(k)} | d_i^{true}] = \mathbb{E}[\sum_j \sum_{k \neq j} W_{ij} W_{ik} | d_i^{true}] \mathbf{C}_{m-1}(d_i^{true}) \tag{B.28}$$

Next,

$$\mathbb{E}[\frac{1}{N} (\mathbf{W}\mathbf{x})' \mathbf{W}^m \mathbf{x}] = \frac{1}{N} \sum_i \mathbb{E}[\mathbf{W}\mathbf{x}_{(i)} (\mathbf{W}^m \mathbf{x})'_{(i)}] = \mathbb{E}[\mathbf{W}\mathbf{x}_{(i)} (\mathbf{W}^m \mathbf{x})'_{(i)}] \tag{B.29}$$

$$= \mathbb{E}[\sum_j \sum_k W_{ij} W_{ik} \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(k)}] \tag{B.30}$$

where the last substitution applies the facts that $(\mathbf{W}\mathbf{x})_{(i)} = \sum_j W_{ij} \mathbf{x}_{(j)}$ and $(\mathbf{W}^m \mathbf{x})_{(i)} = \sum_k W_{ik} \mathbf{W}^{m-1} \mathbf{x}_{(k)}$. Decompose (B.30) into pairs where $k = j$ and $k \neq j$, then apply iterated

expectations.

$$\mathbb{E}[\frac{1}{N}(\mathbf{W}\mathbf{x})'\mathbf{W}^m\mathbf{x}] = \mathbb{E}[\mathbb{E}[\sum_j W_{ij}^2\mathbf{x}_{(i)}(\mathbf{W}^{m-1}\mathbf{x})'_{(i)}|d_i^{true}]] + \mathbb{E}[\mathbb{E}[\sum_j\sum_{k\neq j} W_{ij}W_{ik}\mathbf{x}_{(i)}(\mathbf{W}^{m-1}\mathbf{x})'_{(i)}|d_i^{true}]] \tag{B.31}$$

Substitute (B.24) and (B.28) into (B.31) and simplify, which gives

$$\mathbb{E}[\frac{1}{N}(\mathbf{W}\mathbf{x})'\mathbf{W}^m\mathbf{x}] = \mathbb{E}[\mathbb{E}[\sum_j W_{ij}^2|d_i^{true}]\mathbf{B}_{m-1}(d_i^{true})] + \mathbb{E}[\mathbb{E}[\sum_j\sum_{k\neq j} W_{ij}W_{ik}|d_i^{true}]\mathbf{C}_{m-1}(d_i^{true})] \tag{B.32}$$

Further application of iterated expectations then yields

$$\mathbb{E}[\frac{1}{N}(\mathbf{W}\mathbf{x})'\mathbf{W}^m\mathbf{x}] = \mathbb{E}[\sum_j W_{ij}^2]\mathbf{B}_{m-1}(d_i^{true}) + \mathbb{E}[\sum_j\sum_{k\neq j} W_{ij}W_{ik}]\mathbf{C}_{m-1}(d_i^{true}) \tag{B.33}$$

When $\mathbf{W} = \mathbf{G}$, $\sum_j W_{ij}^2 = \frac{1}{d_i^{true}}$, while $\sum_j\sum_{k\neq j} W_{ij}W_{ik} = 1 - \frac{1}{d_i^{true}}$. Substitution into (B.33) then gives the result in part [1].

When $\mathbf{W} = \mathbf{L}$, $\sum_j W_{ij}^2 = d_i^{true}$, while $\sum_j\sum_{k\neq j} W_{ij}W_{ik} = d_i^{true}(d_i^{true} - 1)$. Substitution into (B.33) then gives the result in part [4].

For Parts [2] and [5], we derive an expression as a function of $\mathbf{W}$ and $\mathbf{W}^*$, then substitute the definitions of $\mathbf{W}$ and $\mathbf{W}^*$ at the end.

Since $W_{ij}(0, d_i^{true}) = 0$ for any d_i^{true} and $W_{ij}^*(0, d_i^{obs})$ for any d_i^{obs},

$$\mathbb{E}[W_{ij}^*W_{ij}|d_i^{true}, d_i^{obs}] = Pr(M_{ij} = 1, L_{ij} = 1|d_i^{true}, d_i^{obs})W_{ij}^*(M_{ij}, d_i^{obs})W_{ij}(1, d_i^{true}) \tag{B.34}$$

Since $M_{ij}, L_{ij} \in \{0, 1\}$,

$$\begin{aligned}&\mathbb{E}[W_{ij}^*W_{ij}\mathbf{x}_{(j)}(\mathbf{W}^{m-1}\mathbf{x})'_{(j)}|d_i^{true}, d_i^{obs}] \\ &\quad = Pr(L_{ij} = 1|d_i^{true})(W_{ij}(1, d_i^{true}))^2\mathbb{E}[\mathbf{x}_{(j)}(\mathbf{W}^{m-1}\mathbf{x})'_{(j)}|M_{ij} = 1, L_{ij} = 1, d_i^{true}, d_i^{obs}]\end{aligned} \tag{B.35}$$

Assumption 5, part [2] and Line (B.34) then imply that this can be simplified as

$$\mathbb{E}[W_{ij}^* W_{ij} \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(j)} | d_i^{true}, d_i^{obs}] = \mathbb{E}[W_{ij}^* W_{ij} | d_i^{true}, d_i^{obs}] \mathbf{B}_{m-1}(d_i^{true}) \tag{B.36}$$

In turn,

$$\mathbb{E}[W_{ij}^* W_{ij} \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(j)} | d_i^{true}] = \left(\sum_d Pr(d_i^{obs} = d | d_i^{true}) \mathbb{E}[W_{ij}^* W_{ij} | d_i^{true}, d_i^{obs}] \right) \mathbf{B}_{m-1}(d_i^{true}) \tag{B.37}$$

$$= \mathbb{E}[W_{ij}^* W_{ij} | d_i^{true}] \mathbf{B}_{m-1}(d_i^{true}) \tag{B.38}$$

Summing up for all $j \neq i$,

$$\mathbb{E}[\sum_j W_{ij}^* W_{ij} \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(j)} | d_i^{true}] = \mathbb{E}[\sum_j W_{ij}^* W_{ij} | d_i^{true}] \mathbf{B}_{m-1}(d_i^{true}) \tag{B.39}$$

And therefore,

$$\mathbb{E}[\sum_j W_{ij}^* W_{ij} \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x}_{(j)})' | d_i^{true}, d_i^{obs}] = \mathbb{E}[\sum_j W_{ij}^* W_{ij} | d_i^{true}, d_i^{obs}] \mathbf{B}_{m-1}(d_i^{true}) \tag{B.40}$$

Similarly,

$$\mathbb{E}[W_{ij}^* W_{ik} | d_i^{true}, d_i^{obs}] = Pr(M_{ij} = 1, L_{ik} = 1 | d_i^{true}, d_i^{obs}) W_{ij}^*(1, d_i^{obs}) W_{ik}(1, d_i^{true}) \tag{B.41}$$

Since $M_{ij}, L_{ik} \in \{0, 1\}$, whenever $k \neq j$ and $W_{ij}(0, d_i^{true}) = W_{ij}^*(0, d_i^{obs}) = 0$,

$$\mathbb{E}[W_{ij}^* W_{ik} \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(k)} | d_i^{true}, d_i^{obs}] =$$
$$= Pr(M_{ij} = 1, L_{ik} = 1 | d_i^{true}, d_i^{obs}) W_{ij}^*(1, d_i^{obs}) W_{ik}(1, d_i^{true}) \mathbb{E}[\mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(k)} | M_{ij} = 1, L_{ik} = 1, d_i^{true}] \tag{B.42}$$

Assumption 5, part [3] and Line (B.41) then imply that this can be simplified as

$$\mathbb{E}[W_{ij}^* W_{ik} \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(k)} | d_i^{true}, d_i^{obs}] = \mathbb{E}[W_{ij}^* W_{ik} | d_i^{true}, d_i^{obs}] \mathbf{C}_{m-1}(d_i^{true}) \tag{B.43}$$

In turn,

$$\mathbb{E}[W_{ij}^* W_{ik} \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(k)} | d_i^{true}] = \left(\sum_d Pr(d_i^{obs} = d | d_i^{true}) \mathbb{E}[W_{ij}^* W_{ik} | d_i^{true}, d_i^{obs}] \right) \mathbf{C}_{m-1}(d_i^{true}) \tag{B.44}$$

$$= \mathbb{E}[W_{ij}^* W_{ik} | d_i^{true}] \mathbf{C}_{m-1}(d_i^{true}) \tag{B.45}$$

Summing up for all $j, k \neq j$,

$$\mathbb{E}[\sum_j \sum_{k \neq j} W_{ij}^* W_{ik} \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(k)} | d_i^{true}] = \mathbb{E}[\sum_j \sum_{k \neq j} W_{ij}^* W_{ik} | d_i^{true}] \mathbf{C}_{m-1}(d_i^{true}) \tag{B.46}$$

Next,

$$\mathbb{E}[\frac{1}{N} (\mathbf{W}^*\mathbf{x})' \mathbf{W}^m \mathbf{x}] = \frac{1}{N} \sum_i \mathbb{E}[\mathbf{W}^* \mathbf{x}_{(i)} (\mathbf{W}^m \mathbf{x})'_{(i)}] = \mathbb{E}[\mathbf{W}^* \mathbf{x}_{(i)} (\mathbf{W}^m \mathbf{x})'_{(i)}] \tag{B.47}$$

$$= \mathbb{E}[\sum_j \sum_k W_{ij}^* W_{ik} \mathbf{x}_{(j)} (\mathbf{W}^{m-1}\mathbf{x})'_{(k)}] \tag{B.48}$$

where the last substitution applies the facts that $(\mathbf{W}^*\mathbf{x})_{(i)} = \sum_j W_{ij}^* \mathbf{x}_{(j)}$ and $(\mathbf{W}^m\mathbf{x})_{(i)} = \sum_k W_{ik} \mathbf{W}^{m-1} \mathbf{x}_{(k)}$. Decompose (B.48) into pairs where $k = j$ and $k \neq j$, then apply iterated expectations.

$$\mathbb{E}[\frac{1}{N} (\mathbf{W}^*\mathbf{x})' \mathbf{W}^m \mathbf{x}] = \mathbb{E}[\mathbb{E}[\sum_j W_{ij}^* W_{ij} \mathbf{x}_{(i)} (\mathbf{W}^{m-1}\mathbf{x})'_{(i)} | d_i^{true}, d_i^{obs}]]$$
$$+ \mathbb{E}[\mathbb{E}[\sum_j \sum_{k \neq j} W_{ij}^* W_{ik} \mathbf{x}_{(i)} (\mathbf{W}^{m-1}\mathbf{x})'_{(i)} | d_i^{true}, d_i^{obs}]] \tag{B.49}$$

Substitute (B.40) and (B.46) into (B.49) and simplify, which gives

$$\mathbb{E}[\frac{1}{N}(\mathbf{W}^*\mathbf{x})'\mathbf{W}^m\mathbf{x}] = \mathbb{E}[\mathbb{E}[\sum_j W_{ij}^* W_{ij} | d_i^{true}]\mathbf{B}_{m-1}(d_i^{true})] + \mathbb{E}[\mathbb{E}[\sum_j \sum_{k\neq j} W_{ij}^* W_{ik} | d_i^{true}]\mathbf{C}_{m-1}(d_i^{true})] \tag{B.50}$$

Further application of iterated expectations then yields

$$\mathbb{E}[\frac{1}{N}(\mathbf{W}^*\mathbf{x})'\mathbf{W}^m\mathbf{x}] = \mathbb{E}[\sum_j W_{ij}^* W_{ij}]\mathbf{B}_{m-1} + \mathbb{E}[\sum_j \sum_{k\neq j} W_{ij}^* W_{ik}]\mathbf{C}_{m-1} \tag{B.51}$$

When $\mathbf{W} = \mathbf{G}$, $\sum_j W_{ij}^* W_{ij} = \frac{1}{d_i^{true}}$, while $\sum_j \sum_{k\neq j} W_{ij} W_{ik} = 1 - \frac{1}{d_i^{true}}$. Substitution into (B.51) then gives the result in part [2].

When $\mathbf{W} = \mathbf{L}$, $\sum_j W_{ij}^* W_{ij} = d_i^{obs}$, while $\sum_j \sum_{k\neq j} W_{ij}^* W_{ij} = d_i^{obs}(d_i^{true} - 1)$. Substitution into (B.51) then gives the result in part [5].

For Parts [3] and [6], we derive an expression as a function of $\mathbf{W}^*$, then substitute the definition of $\mathbf{W}^*$ at the end.

Since $W_{ij}^*(0, d_i^{obs}) = 0$ for any d_i^{obs},

$$\mathbb{E}[(W_{ij}^*)^2 | d_i^{true}, d_i^{obs}] = Pr(M_{ij} = 1, L_{ij} = 1 | d_i^{true}, d_i^{obs})(W_{ij}^*(1, d_i^{obs}))^2 \tag{B.52}$$

Since $M_{ij} \in \{0, 1\}$ and $W_{ij}^*(0, d_i^{obs}) = 0$,

$$\mathbb{E}[(W_{ij}^*)^2 \mathbf{x}_{(j)}(\mathbf{W}^{m-1}\mathbf{x})'_{(j)} | d_i^{obs}] = Pr(M_{ij} = 1, L_{ij} = 1 | d_i^{true}, d_i^{obs})(W_{ij}^*(1, d_i^{obs}))^2 \\ \mathbb{E}[\mathbf{x}_{(j)}(\mathbf{W}^{m-1}\mathbf{x})'_{(j)} | M_{ij} = 1, L_{ij} = 1, d_i^{true}, d_i^{obs}] \tag{B.53}$$

Assumption 5, part [2] and Line (B.52) then imply that this can be simplified as

$$\mathbb{E}[(W_{ij}^*)^2 \mathbf{x}_{(j)}\mathbf{x}'_{(j)} | d_i^{true}, d_i^{obs}] = \mathbb{E}[(W_{ij}^*)^2 | d_i^{true}, d_i^{obs}]\mathbf{B}_0(d_i^{true}) \tag{B.54}$$

In turn,

$$\mathbb{E}[(W_{ij}^*)^2\mathbf{x}_{(j)}\mathbf{x}'_{(j)}|d_i^{true}] = \left(\sum_d Pr(d_i^{obs}=d|d_i^{true})\mathbb{E}[(W_{ij}^*)^2|d_i^{true},d_i^{obs}]\right)\mathbf{B}_0(d_i^{true}) \tag{B.55}$$

$$= \mathbb{E}[(W_{ij}^*)^2|d_i^{true}]\mathbf{B}_0(d_i^{true}) \tag{B.56}$$

Summing up for all $j \neq i$,

$$\mathbb{E}[\sum_j (W_{ij}^*)^2\mathbf{x}_{(j)}\mathbf{x}'_{(j)}|d_i^{true}] = \mathbb{E}[\sum_j (W_{ij}^*)^2|d_i^{true}]\mathbf{B}_0(d_i^{true}) \tag{B.57}$$

Similarly,

$$\mathbb{E}[W_{ij}^*W_{ik}^*|d_i^{obs}] = Pr(M_{ij}=1,M_{ik}=1,L_{ij}=1,L_{ik}=1|d_i^{true},d_i^{obs})W_{ij}(1,d_i^{obs})W_{ik}(1,d_i^{obs}) \tag{B.58}$$

Since $M_{ij}, M_{ik} \in \{0,1\}$, whenever $k \neq j$,

$$\begin{aligned}\mathbb{E}[W_{ij}^*W_{ik}^*\mathbf{x}_{(j)}\mathbf{x}'_{(k)}|d_i^{true},d_i^{obs}] = \\ = Pr(M_{ij}=1,M_{ik}=1,L_{ij}=1,L_{ik}=1|d_i^{true},d_i^{obs},d_i^{true})W_{ij}^*(1,d_i^{obs})W_{ik}^*(1,d_i^{obs}) \\ \mathbb{E}[\mathbf{x}_{(j)}\mathbf{x}'_{(k)}|M_{ij}=1,M_{ik}=1,L_{ij}=1,L_{ik}=1,d_i^{true},d_i^{obs}]\end{aligned} \tag{B.59}$$

Assumption 5, part [3] and Line (B.58) then imply that this can be simplified as

$$\mathbb{E}[W_{ij}^*W_{ik}^*\mathbf{x}_{(j)}\mathbf{x}'_{(k)}|d_i^{true},d_i^{obs}] = \mathbb{E}[W_{ij}^*W_{ik}^*|d_i^{true},d_i^{obs}]\mathbf{C}_0(d_i^{true}) \tag{B.60}$$

In turn,

$$\mathbb{E}[W_{ij}^*W_{ik}^*\mathbf{x}_{(j)}\mathbf{x}'_{(k)}|d_i^{true}] = \left(\sum_d Pr(d_i^{obs}=d|d_i^{true})\mathbb{E}[W_{ij}^*W_{ik}^*|d_i^{true},d_i^{obs}]\right)\mathbf{C}_0(d_i^{true}) \tag{B.61}$$

$$= \mathbb{E}[W_{ij}^*W_{ik}^*|d_i^{true}]\mathbf{C}_0(d_i^{true}) \tag{B.62}$$

Summing up for all $j, k \neq j$,

$$\mathbb{E}[\sum_j \sum_{k \neq j} W_{ij}^* W_{ik}^* \mathbf{x}_{(j)} \mathbf{x}'_{(k)} | d_i^{true}] = \mathbb{E}[\sum_j \sum_{k \neq j} W_{ij}^* W_{ik}^* | d_i^{true}] \mathbf{C}_0(d_i^{true}) \tag{B.63}$$

Next,

$$\mathbb{E}[\frac{1}{N}(\mathbf{W}^*\mathbf{x})'\mathbf{W}^*\mathbf{x}] = \frac{1}{N}\sum_i \mathbb{E}[(\mathbf{W}^*\mathbf{x})_{(i)}(\mathbf{W}^*\mathbf{x})'_{(i)}] = \mathbb{E}[(\mathbf{W}^*\mathbf{x})_{(i)}(\mathbf{W}^*\mathbf{x})'_{(i)}] \tag{B.64}$$

$$= \mathbb{E}[\sum_j \sum_k W_{ij}^* W_{ik}^* \mathbf{x}_{(j)} \mathbf{x}'_{(k)}] \tag{B.65}$$

where the last substitution applies the fact that $(\mathbf{W}^*\mathbf{x})_{(i)} = \sum_j W_{ij}^* \mathbf{x}_{(j)}$. Decompose (B.65) into pairs where $k = j$ and $k \neq j$, then apply iterated expectations.

$$\mathbb{E}[\frac{1}{N}(\mathbf{W}^*\mathbf{x})'\mathbf{W}^*\mathbf{x}] = \mathbb{E}[\mathbb{E}[\sum_j (W_{ij}^*)^2 \mathbf{x}_{(j)} \mathbf{x}'_{(j)} | d_i^{true}]] + \mathbb{E}[\mathbb{E}[\sum_j \sum_{k \neq j} W_{ij}^* W_{ik}^* \mathbf{x}_{(j)} \mathbf{x}'_{(k)} | d_i^{true}]] \tag{B.66}$$

Substitute (B.57) and (B.63) into (B.66) and simplify, which gives

$$\mathbb{E}[\frac{1}{N}(\mathbf{W}^*\mathbf{x})'\mathbf{W}^*\mathbf{x}] = \mathbb{E}[\mathbb{E}[\sum_j (W_{ij}^*)^2 | d_i^{obs}]] \mathbf{B}_0(d_i^{true}) + \mathbb{E}[\mathbb{E}[\sum_j \sum_{k \neq j} W_{ij}^* W_{ik}^* | d_i^{obs}]] \mathbf{C}_0(d_i^{true}) \tag{B.67}$$

Further application of iterated expectations then yields

$$\mathbb{E}[\frac{1}{N}(\mathbf{W}^*\mathbf{x})'\mathbf{W}^*\mathbf{x}] = \mathbb{E}[\sum_j (W_{ij}^*)^2] \mathbf{B}_0(d_i^{true}) + \mathbb{E}[\sum_j \sum_{k \neq j} W_{ij}^* W_{ik}^*] \mathbf{C}_0(d_i^{true}) \tag{B.68}$$

When $\mathbf{W} = \mathbf{G}$, $\sum_j (W_{ij}^*)^2 = \frac{1}{d_i^{obs}}$, while $\sum_j \sum_{k \neq j} W_{ij}^* W_{ik}^* = 1 - \frac{1}{d_i^{obs}}$. Substitution into (B.68) then gives the result in part [3].

When $\mathbf{W} = \mathbf{L}$, $\sum_j (W_{ij}^*)^2 = d_i^{obs}$, while $\sum_j \sum_{k \neq j} W_{ij}^* W_{ik}^* = d_i^{obs}(d_i^{obs} - 1)$. Substitution into (B.68) then gives the result in part [6].

Lemma 3

Preliminaries. For all of Parts [1]-[6], Part [3] of Assumption 1 implies that $(\mathbf{I}-\beta_1\mathbf{W})$ is invertible and thus $\mathbf{y} = (\mathbf{I} - \beta_1\mathbf{W})^{-1}(\mathbf{x}\beta_2 + \mathbf{W}\mathbf{x}\beta_3 + \epsilon)$. Part [3] of Assumption 1 further implies that we can apply the Neumann expansion: $(\mathbf{I} - \beta_1\mathbf{W})^{-1} = \sum_{m=0}^{\infty} \beta_1^m \mathbf{W}^m$. Therefore,

$$\begin{aligned}\mathbf{y} &= \sum_{m=0}^{\infty} \beta_1^m \mathbf{W}^m(\mathbf{x}\beta_2 + \mathbf{W}\mathbf{x}\beta_3) + \sum_{m=0}^{\infty} \beta_1^m \mathbf{W}^m \epsilon \\ &= \mathbf{x}\beta_2 + (\sum_{m=0}^{\infty} \beta_1^m \mathbf{W}^{m+1}\mathbf{x})(\beta_1\beta_2 + \beta_3) + \sum_{m=0}^{\infty} \beta_1^m \mathbf{W}^m \epsilon \end{aligned} \tag{B.69}$$

Parts [1] and [4]. Starting with the expansion in (B.69),

$$\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] = \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}]\beta_2 + (\sum_{m=0}^{\infty} \beta_1^m \mathbb{E}[\mathbf{x}'(\mathbf{W}^{m+1}\mathbf{x})])(\beta_1\beta_2 + \beta_3) + \sum_{m=0}^{\infty} \beta_1^m \mathbb{E}[\mathbf{x}'\mathbf{W}^m\epsilon] \tag{B.70}$$

Assumption 3 implies that the last term is 0. By definition, $\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] = \mathbf{E}_{\mathbf{xx}}$. Therefore,

$$\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] = \mathbf{E}_{\mathbf{xx}}\beta_2 + (\sum_{m=0}^{\infty} \beta_1^m \mathbb{E}[\mathbf{x}'(\mathbf{W}^{m+1}\mathbf{x})])(\beta_1\beta_2 + \beta_3) \tag{B.71}$$

When $\mathbf{W} = \mathbf{G}$, substitute in the expressions from Lemma 1, part [1], which gives the result

$$\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] = \mathbf{E}_{\mathbf{xx}}\beta_2 + (\sum_{m=0}^{\infty} \beta_1^m \mathbb{E}[\mathbf{A}_m(d_i^{true})])(\beta_1\beta_2 + \beta_3) \tag{B.72}$$

When $\mathbf{W} = \mathbf{L}$, substitute in the expressions from Lemma 1, part [3], which gives the result

$$\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] = \mathbf{E}_{\mathbf{xx}}\beta_2 + (\sum_{m=0}^{\infty} \beta_1^m \mathbb{E}[d_i^{true}\mathbf{A}_m(d_i^{true})])(\beta_1\beta_2 + \beta_3) \tag{B.73}$$

Parts [2] and [5]. Starting with the expansion in (B.69),

$$\mathbb{E}[\frac{1}{N}(\mathbf{Wx})'\mathbf{y}] = \mathbb{E}[\frac{1}{N}(\mathbf{Wx})'\mathbf{x}]\beta_2 + (\sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[(\mathbf{Wx})'(\mathbf{W}^{m+1}\mathbf{x})])(\beta_1\beta_2+\beta_3) + \sum_{m=0}^{\infty}\beta_1^m(\mathbf{Wx})'\mathbf{W}^m\epsilon \tag{B.74}$$

Assumption 3 implies that the last term is 0. So,

$$\mathbb{E}[\frac{1}{N}(\mathbf{Wx})'\mathbf{y}] = \mathbb{E}[\frac{1}{N}(\mathbf{Wx})'\mathbf{x}]\beta_2 + (\sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[(\mathbf{Wx})'(\mathbf{W}^{m+1}\mathbf{x})])(\beta_1\beta_2+\beta_3) \tag{B.75}$$

When $\mathbf{W} = \mathbf{G}$, substitute in the expressions from Lemma 1, part [1] and Lemma 2, part [1], which gives the result

$$\mathbb{E}[\frac{1}{N}(\mathbf{Gx})'\mathbf{y}] = \mathbb{E}[\mathbf{A}_0(d_i^{true})]'\beta_2 + (\sum_{m=0}^{\infty}\beta_1^m\left(\mathbb{E}[\mathbf{C}_m(d_i^{true})] + \mathbb{E}[\frac{1}{d^{true}}(\mathbf{B}_m(d_i^{true}) - \mathbf{C}_m(d_i^{true})]\right)(\beta_1\beta_2+\beta_3) \tag{B.76}$$

When $\mathbf{W} = \mathbf{L}$, substitute in the expressions from Lemma 1, part [3] and Lemma 2, part [4], which gives the result

$$\mathbb{E}[\frac{1}{N}(\mathbf{Lx})'\mathbf{y}] = \mathbb{E}\mathbf{A}_0(d_i^{true})]'\beta_2 + (\sum_{m=0}^{\infty}\beta_1^m\left(\mathbb{E}[(d^{true})^2\mathbf{C}_m(d_i^{true})] + \mathbb{E}[d^{true}(\mathbf{B}_m(d_i^{true}) - \mathbf{C}_m(d_i^{true})]\right)(\beta_1\beta_2+\beta \tag{B.77}$$

Parts [3] and [6]. Starting with the expansion in (B.69),

$$\mathbb{E}[\frac{1}{N}(\mathbf{W}^*\mathbf{x})'\mathbf{y}] = \mathbb{E}[\frac{1}{N}(\mathbf{W}^*\mathbf{x})'\mathbf{x}]\beta_2 + (\sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[(\mathbf{W}^*\mathbf{x})'(\mathbf{W}^{m+1}\mathbf{x})])(\beta_1\beta_2+\beta_3) + \sum_{m=0}^{\infty}\beta_1^m(\mathbf{W}^*\mathbf{x})'\mathbf{W}^m\epsilon \tag{B.78}$$

Assumption 3 implies that the last term is 0. So,

$$\mathbb{E}[\frac{1}{N}(\mathbf{W}^*\mathbf{x})'\mathbf{y}] = \mathbb{E}[\frac{1}{N}(\mathbf{W}^*\mathbf{x})'\mathbf{x}]\beta_2 + (\sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[(\mathbf{W}^*\mathbf{x})'(\mathbf{W}^{m+1}\mathbf{x})])(\beta_1\beta_2+\beta_3) \tag{B.79}$$

When $\mathbf{W} = \mathbf{G}$ (and correspondingly $\mathbf{W}^* = \mathbf{H}$), substitute in the expressions from Lemma 1, part [2] and Lemma 2, part [2], which gives the result

$$\mathbb{E}[\frac{1}{N}(\mathbf{Hx})'\mathbf{y}] = \mathbb{E}[\mathbf{A}_0(d_i^{true})]'\beta_2 + \Big(\sum_{m=0}^{\infty}\beta_1^m(\mathbb{E}[\mathbf{C}_m(d_i^{true})] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_m(d_i^{true}) - \mathbf{C}_m(d_i^{true}))]\Big)(\beta_1\beta_2+\beta_3) \tag{B.80}$$

When $\mathbf{W} = \mathbf{L}$ (and correspondingly $\mathbf{W}^* = \mathbf{M}$), substitute in the expressions from Lemma 1, part [4] and Lemma 2, part [5], which gives the result

$$\begin{aligned}\mathbb{E}[\frac{1}{N}(\mathbf{Mx})'\mathbf{y}] = \mathbb{E}[d^{obs}\mathbf{A}_0(d_i^{true})]'\beta_2 + \Big(\sum_{m=0}^{\infty}\beta_1^m\big(&\mathbb{E}[d^{true}d^{obs}\mathbf{C}_m(d_i^{true})] \\ &+ \mathbb{E}[d^{obs}(\mathbf{B}_m(d_i^{true}) - \mathbf{C}_m(d_i^{true}))]\Big)(\beta_1\beta_2+\beta_3)\end{aligned} \tag{B.81}$$

Lemma 4

By the WLLN and Slutzky,

$$\operatorname{plim}\hat{\alpha} = \begin{bmatrix}\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{Gx}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{Gx})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{Gx})'\mathbf{Gx}]\end{bmatrix}^{-1}\begin{bmatrix}\mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{Gx})'\mathbf{y}]\end{bmatrix} \tag{B.82}$$

Substitute $\mathbf{E_{xx}} = \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}]$ and expressions in Lemmas 1 - 3 into Line (B.82).

$$\operatorname{plim}\hat{\alpha} = \begin{bmatrix}\mathbf{E_{xx}} & \mathbb{E}[\mathbf{A}_0] \\ \mathbb{E}[\mathbf{A}_0]' & \mathbb{E}[\mathbf{F}_0]\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{E_{xx}}\beta_2 + (\sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[\mathbf{A}_m])\,(\beta_1\beta_2+\beta_3) \\ \mathbb{E}[\mathbf{A}_0]'\beta_2 + (\sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[\mathbf{F}_m])(\beta_1\beta_2+\beta_3)\end{bmatrix} \tag{B.83}$$

where $\mathbb{E}[\mathbf{F}_m] = \mathbb{E}[\mathbf{C}_m] + \mathbb{E}[\frac{1}{d^{true}}(\mathbf{B}_m - \mathbf{C}_m)]$ for all $m = 0, ..., \infty$ as stated in the lemma statement. Rearrange the expression in Equation (B.83) to show

$$\begin{aligned}
\operatorname{plim} \hat{\alpha} = & \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\mathbf{A}_0] \\ \mathbb{E}[\mathbf{A}_0]' & \mathbb{E}[\mathbf{F}_0] \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\mathbf{A}_0] \\ \mathbb{E}[\mathbf{A}_0]' & \mathbb{E}[\mathbf{F}_0] \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_1\beta_2 + \beta_3 \end{bmatrix} \\
& + \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\mathbf{A}_0] \\ \mathbb{E}[\mathbf{A}_0]' & \mathbb{E}[\mathbf{F}_0] \end{bmatrix}^{-1} \begin{bmatrix} 0 & \sum_{m=1}^{\infty} \beta_1^m \mathbb{E}[\mathbf{A}_m] \\ 0 & \sum_{m=1}^{\infty} \beta_1^m \mathbb{E}[\mathbf{F}_m] \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_1\beta_2 + \beta_3 \end{bmatrix} \qquad \text{(B.84)} \\
= & \begin{bmatrix} \beta_2 \\ \beta_1\beta_2 + \beta_3 \end{bmatrix} + \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\mathbf{A}_0] \\ \mathbb{E}[\mathbf{A}_0]' & \mathbb{E}[\mathbf{F}_0] \end{bmatrix}^{-1} \left(\sum_{m=1}^{\infty} \beta_1^m \begin{bmatrix} \mathbb{E}[\mathbf{A}_m] \\ \mathbb{E}[\mathbf{F}_m] \end{bmatrix} \right) (\beta_1\beta_2 + \beta_3) \qquad \text{(B.85)}
\end{aligned}$$

Lemma 5

By the WLLN and Slutzky,

$$\operatorname{plim} \hat{\alpha} = \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{L}\mathbf{x}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{L}\mathbf{x}] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{y}] \end{bmatrix} \qquad \text{(B.86)}$$

Substitute $\mathbf{E_{xx}} = \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}]$ and expressions in Lemmas 1 - 3 into Line (B.86).

$$\operatorname{plim} \hat{\alpha} = \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[d^{true}\mathbf{A}_0] \\ \mathbb{E}[d^{true}\mathbf{A}_0]' & \mathbb{E}[d^{true}\mathbf{D}_0] \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{E_{xx}} & (\sum_{m=0}^{\infty} \beta_1^m \mathbb{E}[d^{true}\mathbf{A}_m]) \\ \mathbb{E}[d^{true}\mathbf{A}_0]' & (\sum_{m=0}^{\infty} \beta_1^m \mathbb{E}[d^{true}\mathbf{D}_m]) \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_1\beta_2 + \beta_3 \end{bmatrix} \qquad \text{(B.87)}$$

where $\mathbb{E}[d^{true}\mathbf{D}_m] = \mathbb{E}[(d^{true})^2\mathbf{C}_m] + \mathbb{E}[d^{true}(\mathbf{B}_m - \mathbf{C}_m)]$ for all $m = 0, ..., \infty$ as stated in the lemma statement.

Rearrange Equation (B.87) to show

$$\begin{aligned}\operatorname{plim}\hat{\alpha} = &\begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[d^{true}\mathbf{A}_0] \\ \mathbb{E}[d^{true}\mathbf{A}_0]' & \mathbb{E}[d^{true}\mathbf{D}_0] \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[d^{true}\mathbf{A}_0] \\ \mathbb{E}[d^{true}\mathbf{A}_0]' & \mathbb{E}[d^{true}\mathbf{D}_0] \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_1\beta_2+\beta_3 \end{bmatrix} \\ &+ \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[d^{true}\mathbf{A}_0] \\ \mathbb{E}[d^{true}\mathbf{A}_0]' & \mathbb{E}[d^{true}\mathbf{D}_0] \end{bmatrix}^{-1} \begin{bmatrix} 0 & (\sum_{m=1}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m]) \\ 0 & \mathbb{E}[d^{true}](\sum_{m=1}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{D}_m]) \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_1\beta_2+\beta_3 \end{bmatrix} \end{aligned} \tag{B.88}$$

$$= \begin{bmatrix} \beta_2 \\ \beta_1\beta_2+\beta_3 \end{bmatrix} + \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[d^{true}\mathbf{A}_0] \\ \mathbb{E}[d^{true}\mathbf{A}_0]' & \mathbb{E}[d^{true}\mathbf{D}_0] \end{bmatrix}^{-1} \left(\sum_{m=1}^{\infty} \beta_1^m \begin{bmatrix} \mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{true}\mathbf{D}_m] \end{bmatrix} \right) (\beta_1\beta_2+\beta_3) \tag{B.89}$$

Proposition 1

(Note: To simplify notation, we have suppressed the dependence of $\mathbf{A}_m$, $\mathbf{B}_m$, and $\mathbf{C}_m$ on d^{true}. That is, $\mathbf{A}_0 = \mathbf{A}_0(d^{true})$, etc.)

By the WLLN and Slutzky,

$$\operatorname{plim}\hat{\alpha}^{obs} = \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{H}\mathbf{x}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{H}\mathbf{x})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{H}\mathbf{x})'\mathbf{H}\mathbf{x}] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{H}\mathbf{x})'\mathbf{y}] \end{bmatrix} \tag{B.90}$$

By Lemma 3, $\mathbb{E}[\frac{1}{N}(\mathbf{H}\mathbf{x})'\mathbf{y} = \mathbb{E}[\frac{1}{N}(\mathbf{G}\mathbf{x})'\mathbf{y}$. Make this substitution and also multiply by a matrix and its inverse, leading to Equation (B.91).

$$\begin{aligned}\operatorname{plim}\hat{\alpha}^{obs} = &\begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{H}\mathbf{x}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{H}\mathbf{x})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{H}\mathbf{x})'\mathbf{H}\mathbf{x}] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{G}\mathbf{x}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{G}\mathbf{x})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{G}\mathbf{x})'\mathbf{G}\mathbf{x}] \end{bmatrix} \\ &\times \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{G}\mathbf{x}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{G}\mathbf{x})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{G}\mathbf{x})'\mathbf{G}\mathbf{x}] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{G}\mathbf{x})'\mathbf{y}] \end{bmatrix} \end{aligned} \tag{B.91}$$

Note that the final two terms are simply α as defined in Definition 1 and Lemma 4. Make this sibstitution and also substitute $\mathbf{E_{xx}} = \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}]$ and expressions in Lemmas 1-3 into Line (B.91).

$$\operatorname{plim}\hat{\alpha}^{obs} = \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\mathbf{A}_0] \\ \mathbb{E}[\mathbf{A}_0]' & \mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0)] \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\mathbf{A}_0] \\ \mathbb{E}[\mathbf{A}_0]' & \mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{true}}(\mathbf{B}_0 - \mathbf{C}_0)] \end{bmatrix} \alpha \tag{B.92}$$

Further,

$$\begin{aligned} \operatorname{plim}\hat{\alpha}^{obs} &= \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\mathbf{A}_0] \\ \mathbb{E}[\mathbf{A}_0'] & \mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0)] \end{bmatrix}^{-1} \Big(\begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\mathbf{A}_0] \\ \mathbb{E}[\mathbf{A}_0]' & \mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0)] \end{bmatrix} \\ &\quad - \begin{bmatrix} 0 & 0 \\ 0 & \mathbb{E}[(\frac{1}{d^{obs}} - \frac{1}{d^{true}})(\mathbf{B}_0 - \mathbf{C}_0)] \end{bmatrix} \Big)\alpha \qquad \text{(B.93)} \\ &= \alpha - \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\mathbf{A}_0] \\ \mathbb{E}[\mathbf{A}_0'] & \mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0)] \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ 0 & \mathbb{E}[(\frac{1}{d^{obs}} - \frac{1}{d^{true}})(\mathbf{B}_0 - \mathbf{C}_0)] \end{bmatrix} \Big)\alpha \qquad \text{(B.94)} \\ &= \alpha - \begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\mathbf{A}_0] \\ \mathbb{E}[\mathbf{A}_0'] & \mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0)] \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \mathbb{E}[(\frac{1}{d^{obs}} - \frac{1}{d^{true}})(\mathbf{B}_0 - \mathbf{C}_0)] \end{bmatrix} \Big)\alpha_2 \qquad \text{(B.95)} \end{aligned}$$

Next, block matrix inversion gives

$$\begin{bmatrix} \mathbf{E_{xx}} & \mathbb{E}[\mathbf{A}_0] \\ \mathbb{E}[\mathbf{A}_0'] & \mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0)] \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{Z_1}^{-1} & 0 \\ 0 & \mathbf{Z_2}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbb{E}[\mathbf{A}_0](\mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0 \\ -\mathbb{E}[\mathbf{A}_0]'\mathbf{E_{xx}}^{-1} & \mathbf{I} \end{bmatrix} \tag{B.96}$$

where $\mathbf{Z_1} = \mathbf{E_{xx}} - \mathbb{E}[\mathbf{A_0}](\mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0)])^{-1}\mathbb{E}[\mathbf{A}_0]'$, $\mathbf{Z_2} = (\mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0)]) -$

$\mathbb{E}[\mathbf{A}_0]'\mathbf{E_{xx}}^{-1}\mathbb{E}[\mathbf{A}_0]$. Substitute Line (B.96) into (B.95) and simplify, which gives the result

$$\operatorname{plim}\hat{\alpha}^{obs} = \alpha + \begin{bmatrix} \mathbf{Z_1}^{-1}\mathbb{E}[\mathbf{A}_0](\mathbb{E}[\mathbf{C}_0] + \mathbb{E}[\frac{1}{d^{obs}}(\mathbf{B}_0 - \mathbf{C}_0)])^{-1} \\ -\mathbf{Z_2}^{-1} \end{bmatrix} \mathbb{E}[(\frac{1}{d^{obs}} - \frac{1}{d^{true}})(\mathbf{B}_0 - \mathbf{C}_0)]\alpha_2 \tag{B.97}$$

Proposition 2

By the WLLN and Slutzky,

$$\operatorname{plim}\hat{\alpha}^{obs} = \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{M}\mathbf{x}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{M}\mathbf{x}] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{y}] \end{bmatrix} \tag{B.98}$$

Lemma 3 implies

$$\begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{y}] \end{bmatrix} = \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{obs}\mathbf{A}_0'] & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m^{int} \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_1\beta_2 + \beta_3 \end{bmatrix} \tag{B.99}$$

$$\begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{y}] \end{bmatrix} = \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{true}\mathbf{A}_0'] & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_1\beta_2 + \beta_3 \end{bmatrix} \tag{B.100}$$

where $\mathbf{D}_m$ and $\mathbf{D}_m^{int}$ are defined in the proposition statement. Lines (B.99) and (B.100) together imply

$$\begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{y}] \end{bmatrix} = \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{obs}\mathbf{A}_0'] & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m^{int} \end{bmatrix} \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{true}\mathbf{A}_0'] & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{y}] \end{bmatrix} \tag{B.101}$$

Substitute this into Line (B.98) to show

$$\begin{aligned}\operatorname{plim}\hat{\alpha}^{obs} = &\begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{M}\mathbf{x}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{M}\mathbf{x}] \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{E}_{\mathbf{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{obs}\mathbf{A}_0'] & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m^{int} \end{bmatrix} \\ &\times \begin{bmatrix} \mathbf{E}_{\mathbf{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{true}\mathbf{A}_0'] & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{y}] \end{bmatrix}\end{aligned} \tag{B.102}$$

Multiply by a matrix and its inverse, which then yields

$$\begin{aligned}\operatorname{plim}\hat{\alpha}^{obs} = &\begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{M}\mathbf{x}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{M}\mathbf{x})'\mathbf{M}\mathbf{x}] \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{E}_{\mathbf{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{obs}\mathbf{A}_0'] & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m^{int} \end{bmatrix} \\ &\times \begin{bmatrix} \mathbf{E}_{\mathbf{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{true}\mathbf{A}_0'] & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{L}\mathbf{x}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{L}\mathbf{x}] \end{bmatrix} \\ &\times \underbrace{\begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{x}] & \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{L}\mathbf{x}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{x}] & \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{L}\mathbf{x}] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\frac{1}{N}\mathbf{x}'\mathbf{y}] \\ \mathbb{E}[\frac{1}{N}(\mathbf{L}\mathbf{x})'\mathbf{y}] \end{bmatrix}}_{\alpha}\end{aligned} \tag{B.103}$$

Note that the final two terms are simply α. Substitute expressions from Lemmas 1 - 2 to give

$$\begin{aligned}\operatorname{plim}\hat{\alpha}^{obs} = &\underbrace{\begin{bmatrix} \mathbf{E}_{\mathbf{xx}} & \mathbb{E}[d^{obs}\mathbf{A}_0] \\ \mathbb{E}[d^{obs}\mathbf{A}_0]' & \mathbf{D}_0^{obs} \end{bmatrix}^{-1}}_{\mathbf{P}_1^{-1}} \underbrace{\begin{bmatrix} \mathbf{E}_{\mathbf{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{obs}\mathbf{A}_0'] & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m^{int} \end{bmatrix}}_{\mathbf{P}_2} \\ &\times \underbrace{\begin{bmatrix} \mathbf{E}_{\mathbf{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{true}\mathbf{A}_0'] & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m \end{bmatrix}^{-1}}_{\mathbf{P}_3^{-1}} \underbrace{\begin{bmatrix} \mathbf{E}_{\mathbf{xx}} & \mathbb{E}[d^{true}\mathbf{A}_0] \\ \mathbb{E}[d^{true}\mathbf{A}_0]' & \mathbf{D}_0 \end{bmatrix}}_{\mathbf{P}_4}\alpha\end{aligned} \tag{B.104}$$

Define the matrices $\mathbf{P}_1$ - $\mathbf{P}_4$ as shown. So,

$$\operatorname{plim}\hat{\alpha}^{obs} = \mathbf{P}_1^{-1}\mathbf{P}_2\mathbf{P}_3^{-1}\mathbf{P}_4\alpha \tag{B.105}$$

Next, decompose (B.105) as follows:

$$\operatorname{plim}\hat{\alpha}^{obs} = \alpha + \mathbf{P}_1^{-1}(\mathbf{P}_2 - \mathbf{P}_1)\alpha + \mathbf{P}_1^{-1}(\mathbf{P}_4 - \mathbf{P}_3)\alpha + \mathbf{P}_1^{-1}(\mathbf{P}_2 - \mathbf{P}_3)\mathbf{P}_3^{-1}(\mathbf{P}_4 - \mathbf{P}_3)\alpha \tag{B.106}$$

Next, note that

$$(\mathbf{P}_2 - \mathbf{P}_1) = \begin{bmatrix} 0 & \mathbb{E}[(d^{true} - d^{obs})\mathbf{A}_0] \\ 0 & \mathbf{D}_0^{int} - \mathbf{D}_0^{obs} \end{bmatrix} + \sum_{m=1}^{\infty} \beta_1^m \begin{bmatrix} 0 & \mathbb{E}[d^{true}\mathbf{A}_m] \\ 0 & \mathbf{D}_m^{int} \end{bmatrix} \tag{B.107}$$

$$(\mathbf{P}_4 - \mathbf{P}_3) = -\sum_{m=1}^{\infty} \beta_1^m \begin{bmatrix} 0 & \mathbb{E}[d^{true}\mathbf{A}_m] \\ 0 & \mathbf{D}_m \end{bmatrix} \tag{B.108}$$

$$(\mathbf{P}_2 - \mathbf{P}_3) = -\begin{bmatrix} 0 & 0 \\ \mathbb{E}[(d^{true} - d^{obs})\mathbf{A}_0]' & \sum_{m=0}^{\infty} \beta_1^m(\mathbf{D}_m - \mathbf{D}_m^{int}) \end{bmatrix} \tag{B.109}$$

Substitute into Line (B.105), while replacing $\mathbf{D}_0^{int} - \mathbf{D}_0^{obs} = \mathbb{E}[d^{obs}(d^{true} - d^{obs})\mathbf{C}_0]$, to give

$$\begin{aligned}\operatorname{plim}\hat{\alpha}^{obs} = \alpha + \mathbf{P}_1^{-1}\Big(& \begin{bmatrix} 0 & \mathbb{E}[(d^{true} - d^{obs})\mathbf{A}_0] \\ 0 & \mathbb{E}[d^{obs}(d^{true} - d^{obs})\mathbf{C}_0] \end{bmatrix} - \sum_{m=1}^{\infty} \beta_1^m \begin{bmatrix} 0 & 0 \\ 0 & (\mathbf{D}_m - \mathbf{D}_m^{int}) \end{bmatrix} \\ & + \sum_{m=1}^{\infty} \beta_1^m \begin{bmatrix} 0 & 0 \\ \mathbb{E}[(d^{true} - d^{obs})\mathbf{A}_0]' & \sum_{m=0}^{\infty} \beta_1^m(\mathbf{D}_m - \mathbf{D}_m^{int}) \end{bmatrix} \mathbf{P}_3^{-1} \begin{bmatrix} 0 & \mathbb{E}[d^{true}\mathbf{A}_m] \\ 0 & \mathbf{D}_m \end{bmatrix} \Big)\alpha \end{aligned} \tag{B.110}$$

which simplifies to

$$\begin{aligned}\operatorname{plim}\hat{\alpha}^{obs} = \alpha + \mathbf{P}_1^{-1}\Big(& \begin{bmatrix} \mathbb{E}[(d^{true} - d^{obs})\mathbf{A}_0] \\ \mathbb{E}[d^{obs}(d^{true} - d^{obs})\mathbf{C}_0] \end{bmatrix} - \sum_{m=1}^{\infty} \beta_1^m \begin{bmatrix} 0 \\ (\mathbf{D}_m - \mathbf{D}_m^{int}) \end{bmatrix} \\ & + \sum_{m=1}^{\infty} \beta_1^m \begin{bmatrix} 0 & 0 \\ \mathbb{E}[(d^{true} - d^{obs})\mathbf{A}_0]' & \sum_{m=0}^{\infty} \beta_1^m(\mathbf{D}_m - \mathbf{D}_m^{int}) \end{bmatrix} \mathbf{P}_3^{-1} \begin{bmatrix} \mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbf{D}_m \end{bmatrix} \Big)\alpha_2 \end{aligned} \tag{B.111}$$

Corollary 2

The assumption that the covariate is assigned i.i.d. implies that $\mathbb{E}[d^{true}\mathbf{A}_0] = \mathbb{E}[d^{obs}\mathbf{A}_0] = \mathbb{E}[(d^{true} - d^{obs})\mathbf{A}_0] = 0$. Further, since $\mathbf{C}_0(d^{true}) = 0$ for all d^{true}, it also follows that $\mathbf{D}_0^{int} - \mathbf{D}_0^{obs} = 0$. Therefore, the statement in Proposition 2 simplifies to (with some slight rearranging of summations)

$$\begin{aligned}\operatorname{plim}\hat{\alpha}^{obs} = \alpha + \mathbf{P}_1^{-1}\Bigg(&\begin{bmatrix} 0 \\ -\sum_{m=1}^{\infty}\beta_1^m(\mathbf{D}_m - \mathbf{D}_m^{int}) \end{bmatrix} \\ &+ \begin{bmatrix} 0 & 0 \\ 0 & \sum_{m=0}^{\infty}\beta_1^m(\mathbf{D}_m - \mathbf{D}_m^{int}) \end{bmatrix} \mathbf{P}_3^{-1} \begin{bmatrix} \sum_{m=1}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \sum_{m=1}^{\infty}\beta_1^m\mathbf{D}_m \end{bmatrix} \Bigg)\alpha_2 \end{aligned} \tag{B.112}$$

Define $\mathbf{F} = \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m$, $\mathbf{F}^{int} = \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m^{int}$ and substitute into Line (B.112), which gives

$$\operatorname{plim}\hat{\alpha}^{obs} = \alpha + \mathbf{P}_1^{-1}\Bigg(\begin{bmatrix} 0 \\ (\mathbf{F}^{int} - \mathbf{F}) + (\mathbf{D}_0 - \mathbf{D}_0^{int}) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{F} - \mathbf{F}^{int} \end{bmatrix} \mathbf{P}_3^{-1} \begin{bmatrix} \sum_{m=1}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbf{F} - \mathbf{D}_0 \end{bmatrix} \Bigg)\alpha_2 \tag{B.113}$$

Next, note that

$$\begin{bmatrix} 0 & 0 \\ 0 & \mathbf{F} - \mathbf{F}^{int} \end{bmatrix} \mathbf{P}_3^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{I} - \mathbf{F}^{int}\mathbf{F}^{-1} \end{bmatrix} \tag{B.114}$$

and therefore, plugging this into Equation (B.113),

$$\operatorname{plim}\hat{\alpha}^{obs} = \alpha + \mathbf{P}_1^{-1}\Bigg(\begin{bmatrix} 0 \\ (\mathbf{F}^{int} - \mathbf{F}) + (\mathbf{D}_0 - \mathbf{D}_0^{int}) \end{bmatrix} + \begin{bmatrix} 0 \\ (\mathbf{I} - \mathbf{F}^{int}\mathbf{F}^{-1})(\mathbf{F} - \mathbf{D}_0) \end{bmatrix} \Bigg)\alpha_2 \tag{B.115}$$

which in turn simplifies to

$$\operatorname{plim} \hat{\alpha}^{obs} = \alpha + \mathbf{P}_1^{-1} \begin{bmatrix} 0 \\ \mathbf{F}^{int}\mathbf{F}^{-1}\mathbf{D}_0 - \mathbf{D}_0^{int} \end{bmatrix} \alpha_2 \tag{B.116}$$

$$= \alpha + \begin{bmatrix} 0 \\ (\mathbf{D}_0^{obs})^{-1}\mathbf{F}^{int}\mathbf{F}^{-1}\mathbf{D}_0 - (\mathbf{D}_0^{obs})^{-1}\mathbf{D}_0^{int} \end{bmatrix} \alpha_2 \tag{B.117}$$

Finally, note that the i.i.d. assumption implies that $\mathbf{D}_0 = \mathbb{E}[d^{true}]\mathbf{E_{xx}}$, $\mathbf{D}_0^{int} = \mathbb{E}[d^{obs}]\mathbf{E_{xx}}$, and $\mathbf{D}_0^{obs} = \mathbb{E}[d^{obs}]\mathbf{E_{xx}}$. Substitute these into Equation (B.115) to give the result

$$\operatorname{plim} \hat{\alpha}^{obs} = \alpha + \begin{bmatrix} 0 \\ \frac{\mathbb{E}[d^{true}]}{\mathbb{E}[d^{obs}]}\mathbf{E_{xx}}^{-1}\mathbf{F}^{int}\mathbf{F}^{-1}\mathbf{E_{xx}} - \mathbf{I} \end{bmatrix} \alpha_2 \tag{B.118}$$

Corollary 3

Let $p = \frac{\mathbb{E}[d^{obs}]}{\mathbb{E}[d^{true}]}$. Random missingness implies, among other things, that $\mathbb{E}[d^{obs}\mathbf{A}_0] = p\mathbb{E}[d^{true}\mathbf{A}_0]$ and also that $\mathbf{D}_m^{int} = p\mathbf{D}_m$ for all m. Therefore, using the definitions of $\mathbf{P}_2$ and $\mathbf{P}_3$ in the proof of Proposition 2,

$$\mathbf{P}_2 = \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{obs}\mathbf{A}_0'] & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m^{int} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & p\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{E_{xx}} & \sum_{m=0}^{\infty}\beta_1^m\mathbb{E}[d^{true}\mathbf{A}_m] \\ \mathbb{E}[d^{true}\mathbf{A}_0'] & \sum_{m=0}^{\infty}\beta_1^m\mathbf{D}_m \end{bmatrix} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & p\mathbf{I} \end{bmatrix} \mathbf{P}_3 \tag{B.119}$$

And then the expression in Line (B.105) simplies to

$$\operatorname{plim} \hat{\alpha}^{obs} = \mathbf{P}_1^{-1} \begin{bmatrix} \mathbf{I} & 0 \\ 0 & p\mathbf{I} \end{bmatrix} \mathbf{P}_4\alpha \tag{B.120}$$

Next, note that

$$\begin{aligned}\begin{bmatrix}\mathbf{I} & 0\\ 0 & p\mathbf{I}\end{bmatrix}\mathbf{P}_4 &= \begin{bmatrix}\mathbf{I} & 0\\ 0 & p\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{E_{xx}} & \mathbb{E}[d^{true}\mathbf{A}_0]\\ \mathbb{E}[d^{true}\mathbf{A}_0]' & \mathbf{D}_0^{true}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{E_{xx}} & \mathbb{E}[d^{true}\mathbf{A}_0]\\ \mathbb{E}[d^{obs}\mathbf{A}_0]' & \mathbf{D}_0^{int}\end{bmatrix}\\ &= \underbrace{\begin{bmatrix}\mathbf{E_{xx}} & \mathbb{E}[d^{obs}\mathbf{A}_0]\\ \mathbb{E}[d^{obs}\mathbf{A}_0]' & \mathbf{D}_0^{obs}\end{bmatrix}}_{\mathbf{P}_1} + \begin{bmatrix}0 & \mathbb{E}[(d^{true}-d^{obs})\mathbf{A}_0]\\ 0 & \mathbf{D}_0^{int}-\mathbf{D}_0^{obs}\end{bmatrix}\end{aligned} \tag{B.121}$$

Substitute this into Line (B.120) while replacing $\mathbf{D}_0^{int} - \mathbf{D}_0^{obs} = \mathbb{E}[d^{obs}(d^{true} - d^{obs})\mathbf{C}_0]$ to show

$$\operatorname{plim}\hat{\alpha}^{obs} = \alpha + \mathbf{P}_1^{-1}\begin{bmatrix}\mathbb{E}[(d^{true}-d^{obs})\mathbf{A}_0]\\ \mathbb{E}[d^{obs}(d^{true}-d^{obs})\mathbf{C}_0]\end{bmatrix}\alpha_2 \tag{B.122}$$

With random missingness, $d^{obs}|d^{true}$ follows a $Binomial(d^{true}, p)$ distribution. Therefore, for any d^{true},

$$\begin{aligned}\mathbb{E}[(d^{true}-d^{obs})d^{obs}|d^{true}] &= \mathbb{E}[d^{true}d^{obs}|d^{true}] - \mathbb{E}[(d^{obs})^2|d^{true}]\\ &= p(d^{true})^2 - (\mathbb{V}[(d^{obs})^2|d^{true}] + (\mathbb{E}[d^{obs}|d^{true}])^2)\\ &= p(d^{true})^2 - (p(1-p)d^{true} + p^2(d^{true})^2) = p(1-p)d^{true}(d^{true}-1)\end{aligned} \tag{B.123}$$

Note also that $\mathbb{E}[(d^{true}-d^{obs})\mathbf{A}_0] = (1-p)\mathbb{E}[d^{true}\mathbf{A}_0]$. Make these substitutions into Line (B.122), giving

$$\operatorname{plim}\hat{\alpha}^{obs} = \alpha + (1-p)\mathbf{P}_1^{-1}\begin{bmatrix}\mathbb{E}[d^{true}\mathbf{A}_0]\\ p\mathbb{E}[d^{true}(d^{true}-1)\mathbf{C}_0]\end{bmatrix}\alpha_2 \tag{B.124}$$

Block matrix inversion then provides that

$$\begin{aligned}\mathbf{P}_1^{-1} &= \begin{bmatrix}\mathbf{Z}_1^{-1} & 0\\ 0 & \mathbf{Z}_2^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{I} & -\mathbb{E}[d^{obs}\mathbf{A}_0](\mathbf{D}_0^{obs})^{-1}\\ -\mathbb{E}[d^{obs}\mathbf{A}_0]'{\mathbf{E}_{\mathbf{xx}}}^{-1} & \mathbf{I}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{Z}_1^{-1} & 0\\ 0 & \mathbf{Z}_2^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{I} & -p\mathbb{E}[d^{true}\mathbf{A}_0](\mathbf{D}_0^{obs})^{-1}\\ -p\mathbb{E}[d^{true}\mathbf{A}_0]'{\mathbf{E}_{\mathbf{xx}}}^{-1} & \mathbf{I}\end{bmatrix}\end{aligned} \tag{B.125}$$

and therefore

$$\begin{aligned}\operatorname{plim}\hat{\alpha}^{obs} = \alpha + p(1-p)\begin{bmatrix}\mathbf{Z}_1^{-1} & 0\\ 0 & \mathbf{Z}_2^{-1}\end{bmatrix}\Bigg(&\begin{bmatrix}-\mathbb{E}[d^{true}\mathbf{A}_0](\mathbf{D}_0^{obs})^{-1}\\ \mathbf{I}\end{bmatrix}\mathbb{E}[d^{true}(d^{true}-1)\mathbf{C}_0]\\ &+\begin{bmatrix}\frac{1}{p}\mathbf{I}\\ -\mathbb{E}[d^{true}\mathbf{A}_0]'{\mathbf{E}_{\mathbf{xx}}}^{-1}\end{bmatrix}\mathbb{E}[d^{true}\mathbf{A}_0]\Bigg)\alpha_2\end{aligned} \tag{B.126}$$

Corollary 4

Since $\mathbf{x}_{(j)} = 1$ for all j, it follows that $\mathbf{E_{xx}} = \mathbf{A}_0(d^{true}) = \mathbf{B}_0(d^{true}) = \mathbf{C}_0(d^{true}) = 1$ for all d^{true}. Therefore,

$$\mathbf{D}_0^{obs} = \mathbb{E}[d^{obs}(d^{obs}-1)\mathbf{C}_0] + \mathbb{E}[d^{obs}\mathbf{B}_0] = \mathbb{E}[(d^{obs})^2] \tag{B.127}$$

and in turn

$$\mathbf{Z}_2 = \mathbf{D}_0^{obs} - \mathbb{E}[d^{obs}\mathbf{A}_0]'{\mathbf{E}_{\mathbf{xx}}}^{-1}\mathbb{E}[d^{obs}\mathbf{A}_0] = \mathbb{E}[(d^{obs})^2] - \mathbb{E}[d^{obs}]^2 = \mathbb{V}[d^{obs}] \tag{B.128}$$

Substitute these into the expression for $\operatorname{plim}\hat{\alpha}_2^{obs}$ given in Corollary 3, which yields

$$\begin{aligned}\operatorname{plim}\hat{\alpha}_2^{obs} &= \alpha_2 + p(1-p)\frac{1}{\mathbb{V}[d^{obs}]}\Big(\mathbb{E}[d^{true}(d^{true}-1)] - \mathbb{E}[d^{true}]^2\Big)\alpha_2\\ &= \alpha_2 + p(1-p)\frac{1}{\mathbb{V}[d^{obs}]}(\mathbb{V}[d^{true}] - \mathbb{E}[d^{true}])\alpha_2\end{aligned} \tag{B.129}$$

Since $d^{obs}|d^{true} \sim Binomial(d^{true}, p)$, it follows that

$$\begin{aligned}\mathbb{E}[(d^{obs})^2] &= \mathbb{E}[\mathbb{E}[(d^{obs})^2|d^{true}]] = \mathbb{E}[\mathbb{V}[d^{obs}|d^{true}]] + (\mathbb{E}[d^{obs}|d^{true}])^2] \\ &= \mathbb{E}[p(1-p)d^{true} + p^2(d^{true})^2] = p(1-p)\mathbb{E}[d^{true}] + p^2\mathbb{E}[(d^{true})^2] \quad \text{(B.130)} \\ \mathbb{V}[d^{obs}] &= \mathbb{E}[(d^{obs})^2] - \mathbb{E}[d^{obs}]^2 = p(1-p)\mathbb{E}[d^{true}] + p^2\mathbb{V}[d^{true}] \quad \text{(B.131)}\end{aligned}$$

Substitute into (B.129) and simplify:

$$\operatorname{plim} \hat{\alpha}_2^{obs} = \alpha_2 + \frac{\mathbb{V}[d^{true}] - \mathbb{E}[d^{true}]}{\frac{p}{1-p}\mathbb{V}[d^{true}] + \mathbb{E}[d^{true}]}\alpha_2 \quad \text{(B.132)}$$

Corollary 5

The assumption that $\mathbf{x} = [\iota, \mathbf{T}]$ and $\mathbf{T}$ is i.i.d. and mean zero implies the following:

$$\mathbf{E_{xx}} = \mathbf{B}_0(d^{true}) = \begin{bmatrix} 1 & 0 \\ 0 & \mathbb{V}[\mathbf{T}] \end{bmatrix}, \quad \mathbf{A}_0(d^{true}) = \mathbf{C}_0(d^{true}) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{(B.133)}$$

In turn,

$$\mathbf{D}_0^{obs} = \mathbb{E}[d^{obs}(d^{obs}-1)\mathbf{C}_0] + \mathbb{E}[d^{obs}\mathbf{B}_0] = \begin{bmatrix} \mathbb{E}[(d^{obs})^2] & 0 \\ 0 & \mathbb{E}[d^{obs}]\mathbb{V}[\mathbf{T}] \end{bmatrix} \quad \text{(B.134)}$$

Substitute into the expression in Corollary 3,

$$\operatorname{plim}\hat{\alpha}^{obs} = \alpha + p(1-p)\begin{bmatrix}\mathbf{Z}_1^{-1} & 0\\ 0 & \mathbf{Z}_2^{-1}\end{bmatrix} \Bigg(\begin{bmatrix}-\frac{\mathbb{E}[d^{true}]}{\mathbb{E}[(d^{obs})^2]} & 0\\ 0 & 0\\ 1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\mathbb{E}[d^{true}(d^{true}-1)] & 0\\ 0 & 0\end{bmatrix} + \begin{bmatrix}\frac{1}{p} & 0\\ 0 & \frac{1}{p}\\ -\mathbb{E}[d^{true}] & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\mathbb{E}[d^{true}] & 0\\ 0 & 0\end{bmatrix}\Bigg)\alpha_2 \tag{B.135}$$

which simplifies to

$$\operatorname{plim}\hat{\alpha}^{obs} = \alpha + p(1-p)\begin{bmatrix}\mathbf{Z}_1^{-1} & 0\\ 0 & \mathbf{Z}_2^{-1}\end{bmatrix}\begin{bmatrix}\frac{\mathbb{E}[d^{true}]}{p\mathbb{E}[(d^{obs})^2]}(\mathbb{E}[(d^{obs})^2] - p(\mathbb{E}[(d^{true})^2] - \mathbb{E}[d^{true}]))\\ 0\\ \mathbb{V}[d^{true}] - \mathbb{E}[d^{true}]\\ 0\end{bmatrix}\alpha_{2,1} \tag{B.136}$$

Next, note that, under the assumptions stated,

$$\mathbf{Z}_1 = \begin{bmatrix}1 & 0\\ 0 & \mathbb{V}[\mathbf{T}]\end{bmatrix} - \begin{bmatrix}\mathbb{E}[d^{obs}] & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\mathbb{E}[(d^{obs})^2] & 0\\ 0 & \mathbb{E}[d^{obs}]\mathbb{V}[\mathbf{T}]\end{bmatrix}^{-1}\begin{bmatrix}\mathbb{E}[d^{obs}] & 0\\ 0 & 0\end{bmatrix} = \begin{bmatrix}\frac{\mathbb{V}[d^{obs}]}{\mathbb{E}[(d^{obs})^2]} & 0\\ 0 & \mathbb{V}[\mathbf{T}]\end{bmatrix} \tag{B.137}$$

$$\mathbf{Z}_2 = \begin{bmatrix}\mathbb{E}[(d^{obs})^2] & 0\\ 0 & \mathbb{E}[d^{obs}]\mathbb{V}[\mathbf{T}]\end{bmatrix} - \begin{bmatrix}\mathbb{E}[d^{obs}] & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & \mathbb{V}[\mathbf{T}]\end{bmatrix}^{-1}\begin{bmatrix}\mathbb{E}[d^{obs}] & 0\\ 0 & 0\end{bmatrix} = \begin{bmatrix}\mathbb{V}[d^{obs}] & 0\\ 0 & \mathbb{E}[d^{obs}]\mathbb{V}[\mathbf{T}]\end{bmatrix} \tag{B.138}$$

Additionally, since $d^{obs}|d^{true} \sim Binomial(p, d^{true})$, it follows that

$$\begin{aligned}\mathbb{E}[(d^{obs})^2] &= \mathbb{E}[\mathbb{E}[(d^{obs})^2|d^{true}]] = \mathbb{E}[\mathbb{V}[d^{obs}|d^{true}]] + (\mathbb{E}[d^{obs}|d^{true}])^2] \\ &= \mathbb{E}[p(1-p)d^{true} + p^2(d^{true})^2] = p(1-p)\mathbb{E}[d^{true}] + p^2\mathbb{E}[(d^{true})^2] \end{aligned} \tag{B.139}$$

and therefore,

$$\operatorname{plim} \hat{\alpha}^{obs} = \alpha + \begin{bmatrix} -\frac{\mathbb{E}[d^{true}]\left((1-p)(\mathbb{E}[(d^{true})^2]-\mathbb{E}[d^{true}])-\mathbb{E}[d^{true}]\right)}{\frac{p}{1-p}\mathbb{V}[d^{true}]+\mathbb{E}[d^{true}]} \\ 0 \\ \frac{\mathbb{V}[d^{true}]-\mathbb{E}[d^{true}]}{\frac{p}{1-p}\mathbb{V}[d^{true}]+\mathbb{E}[d^{true}]} \\ 0 \end{bmatrix} \alpha_{2,1} \tag{B.140}$$

B.2 Simulation Study

In this appendix, we illustrate our main results in Sections 2.4 and 2.5 through simulations. Appendix B.2.1 describes the data we use, how we simulate outcomes, and estimation. Appendices B.2.2 and B.2.3 illustrate results for the i.i.d and weak homophily cases, respectively. Finally, Appendix B.2.4 illustrates our results for regressions that include number of links, a special case of the linear-in-sums model as discussed in Subsection 2.5.2 in the main text.

B.2.1 Data Description

Network

We employ the widely-used network data from Banerjee et al. (2012), which contains data from 75 villages in rural India. Through a detailed network survey, the authors collected information on 12 types of network links based on the relationships such as borrowing or lending money between the residents and its households. We use the "ALL" household network which is the union network of all types; i.e. for each village, if a resident of household i has any type of relationship with any resident of household j, then an edge exists between the two households. From this definition, we construct the adjacency matrix of all 75 networks and construct one large block-diagonal symmetric (undirected) adjacency matrix, corresponding to $\mathbf{L}$ as defined in the main text. The total number of nodes (all households) of this network is $14,904$. Mean degree is 8.9723, while its standard deviation is 7.2779.

Data Generating Process

We simulate the data generating process as described in Section 2.2. Namely, the outcome $\mathbf{y}$ in reduced form is given in Equation (B.141).

$$\mathbf{y} = (\mathbf{I} - \beta_1 \mathbf{W})^{-1} (\mathbf{x}\beta_2 + \mathbf{W}\mathbf{x}\beta_3 + \epsilon) \tag{B.141}$$

In Equation (B.141), $\mathbf{W} = \mathbf{G}$ for the linear-in-means model and $\mathbf{W} = \mathbf{L}$ for the linear-in-sums model. Note that outcomes are simulated according to the "true" $\mathbf{W}$.

We construct the covariate $\mathbf{x}$ two different ways. In the i.i.d. cases, each element of $\mathbf{x}$ is drawn from $N(0, 1)$ randomly. Alternatively, in the weak homophily case, we generate $\mathbf{x}$ through the following process:

$$\mathbf{x} = \mathbf{G}\mathbf{v} + \mathbf{u}, \text{ where } \begin{bmatrix} v_i \\ u_i \end{bmatrix} \sim_{i.i.d.} N(0, \begin{bmatrix} \mathbf{I}_d & 0 \\ 0 & 0.1\mathbf{I}_d \end{bmatrix}) \tag{B.142}$$

That is, $\mathbf{v}$ and $\mathbf{u}$ are $N \times d$ matrices with each element independently drawn. We then construct $\mathbf{x}$ in a way that generates correlation among connected agents ($\mathbf{x} = \mathbf{G}\mathbf{v} + \mathbf{u}$).

In all simulations, the non-random, "structural" parameters are fixed as follows:

$$\beta_1 = \begin{cases} 0.1 & \text{Linear-in-Means} \\ 0.01 & \text{Linear-in-Sums} \end{cases}$$

$$\beta_2 = \iota_d$$

$$\beta_3 = 0.5\iota_d$$

where ι_d is a d-dimensional vector of 1's. [1] Finally, each element in ϵ in Equation (B.141) is drawn independently from $N(0, 1)$ for all types and models. With the exception of the simulations in Appendix B.2.4, we set $d = 1$ for simplicity.

Missingness Mechanisms

We simulate two missingness mechanisms to construct the observed networks $\mathbf{H}$ and $\mathbf{M}$. These are two special cases that are discussed in Sections 2.4 - 2.6. For *random missingness*, each edge

[1] Note that large β_1 in the linear-in-sums model leads to noisiness in inverting $\mathbf{I} - \beta_1\mathbf{W}$. The limit of β_1 causing extreme cases differs by the network data. In our simulations, using values exceeding 0.03 caused extremely high variance of generated outcome values $\mathbf{y}$. We also ran simulations with various values of parameters but the results remained qualitatively unchanged.

from the true network is observed with probability $p \in (0, 1]$. In practice, we use p from 0.1 to 1 in increments of 0.1. In contrast, *censored missingness* is simulated by removing edges of agent i if the agent's degree exceeds a certain threshold. That is, for each agent i, if degree does not exceed the threshold, we do not remove any edges. If degree does exceed the threshold, we select a subset of the nodes without replacement, such that d_i^{obs} is equal to the maximum number of links allowed by the censoring rule. The thresholds in our results are selected by deciles of the degree distribution. From 0.1 to 1, these values are $(1, 3, 4, 6, 7, 9, 12, 15, 19, 90)$.

Once we employ each mechanism, we construct the adjacency matrix $\mathbf{M}$. For linear-in-means specifications, we normalize each row of $\mathbf{M}$ by in-degree to construct $\mathbf{H}$.[2]

Estimation within Simulations

We repeat the above process across a large number of simulations, indexed $(s) = 1, ..., S$. For each simulation draw (s), we generate the data $(\mathbf{y}_{(s)}, \mathbf{x}_{(s)})$ as described above, and we also simulate $\mathbf{W}_{(s)}$ and $\mathbf{W}^*_{(s)}$ as appropriate. For each (s), we then estimate the *observed-data RF Estimator* as defined in Definition 4 and given in Equation (B.143).

$$\hat{\alpha}^{obs}_{(s)} = \left(\left[\mathbf{x}_{(s)}, \mathbf{W}^*_{(s)}\mathbf{x}_{(s)} \right]' \left[\mathbf{x}_{(s)}, \mathbf{W}^*_{(s)}\mathbf{x}_{(s)} \right] \right)^{-1} \left[\mathbf{x}_{(s)}, \mathbf{W}^*_{(s)}\mathbf{x} \right]' \mathbf{y}_{(s)} \tag{B.143}$$

In Equation (B.143), $\mathbf{W}^*_{(s)} = \mathbf{H}_{(s)}$ for the linear-in-means model, while $\mathbf{W}^*_{(s)} = \mathbf{M}_{(s)}$ for the linear-in-sums model.

B.2.2 I.I.D.

Linear-in-Means Model

Figure B.1 summarizes the simulated estimates of $\hat{\alpha}^{obs}_{(s)} = (\hat{\alpha}^{obs}_{(s)1}, \hat{\alpha}^{obs}_{(s)2})$. The first column (Panels (a) and (c)) shows the results using random missingness where in each plot horizon-

[2]Note that if the true network is undirected, this procedure may result in constructing a directed network. For purposes of these simulations, we impose no restriction on the symmetry of $\mathbf{M}$ or $\mathbf{H}$. Results are qualitatively identical (in terms of direction of inconsistency) if we alternatively construct undirected networks through unions or intersections of the edges that remain after employing the appropriate missingness procedure.

tal bars denote ± 2 standard deviations around the mean of the simulated estimates at each $p \in \{0.1, 0.2, ..., 1\}$. The second column (Panels (b) and (d)) describes the results of the estimates using censored missingness, where the ten bars correspond to the two sigma interval at each degree quantile for $(0.1, 0.2, ..., 1)$. The red dashed line gives true α_1 or α_2, corresponding to the mean value of the simulated estimates with the true network, which are by definition consistent (see Definition 5).

The results below align with what is shown in Equation (2.14) of the main text. Regardless of the missingness mechanism, $\hat{\alpha}_1^{obs}$ is a consistent estimator of α_1, as shown in Panels (a) and (b). In contrast, as shown in Panels (c) and (d), $\hat{\alpha}_2^{obs}$ gives attenuated estimates, since $\mathbb{E}[1/d^{obs}] \geq \mathbb{E}[1/d^{true}]$ always holds; further, the results are more attenuated when fewer links are observed, a pattern that is clear from Figure B.1, Panels (c)-(d).

Linear-in-Sums Model

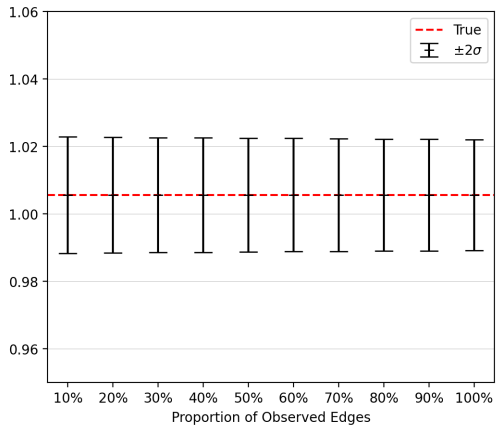
Figure B.2 simmarizes the analogous results for the linear-in-sums case. We see that the results are in line with the descriptions in Remarks 3, 4, and 5 from the main paper. The simulated estimates for $\hat{\alpha}_1^{obs}$ are consistent for both types of missingness (Panels (a) and (b)). In contrast, the estimates for $\hat{\alpha}_2^{obs}$ are consistent when links are missing randomly (Panel (c)), but attenuated under censored missingness (Panel (d)).

B.2.3 Weak Homophily

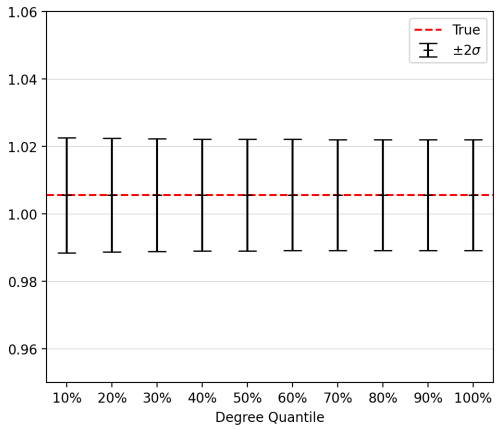
Linear-in-Means Model

Next, we show results under weak homophily, as defined in Assumption 7. Figure B.3 summarizes results under this version of the linear-in-means data generating process. We can see that the results are in line with Remarks 1 and 2 in Section 2.5, which state that the sign of the inconsistency of $\hat{\alpha}_1$ is the same as the sign of $\mathbb{E}[\mathbf{A}_0]\alpha_2$ (positive, given the assumed DGP) and that estimates of $\hat{\alpha}_2$ are attenuated. Note that these results both hold regardless of the missingness mechanism (compare Panels (a) and (c) to (b) and (d), respectively).

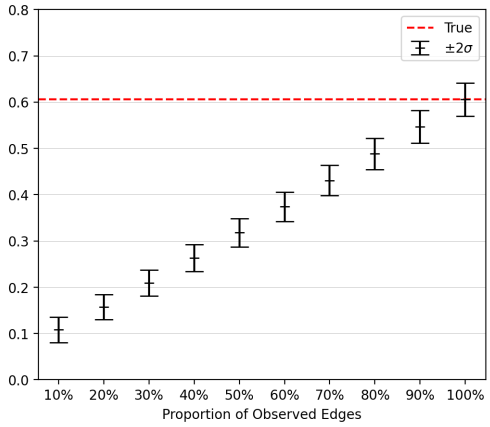
Figure B.1: Simulated Estimates under i.i.d. Assumption
(Linear-in-Means Model)



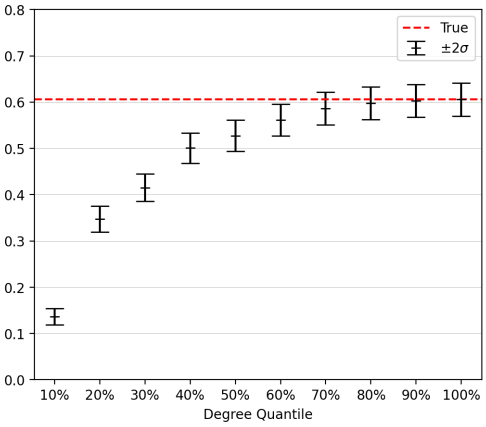
a Estimates for $\hat{\alpha}_1^{obs,sim}$
(Random Missingness)



b Estimates for $\hat{\alpha}_1^{obs,sim}$
(Censored Missingness)



c Estimates for $\hat{\alpha}_2^{obs,sim}$
(Random Missingness)

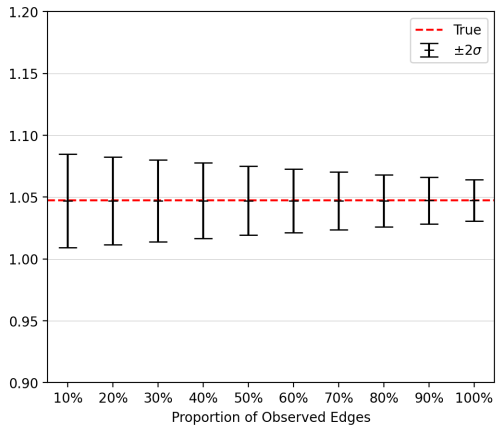


d Estimates for $\hat{\alpha}_2^{obs,sim}$
(Censored Missingness)

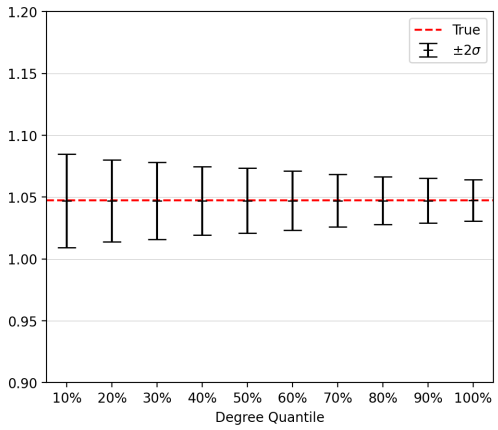
Linear-in-Sums Model

In contrast, Figure B.4 summarizes the simulated estimates of the linear-in-sums model under weak homophily. We see that the results are in line with 4. As theory predicts, the inconsistency in the simulated estimates for $\hat{\alpha}_1^{obs}$ have the same sign as true α_2 for random missingness and

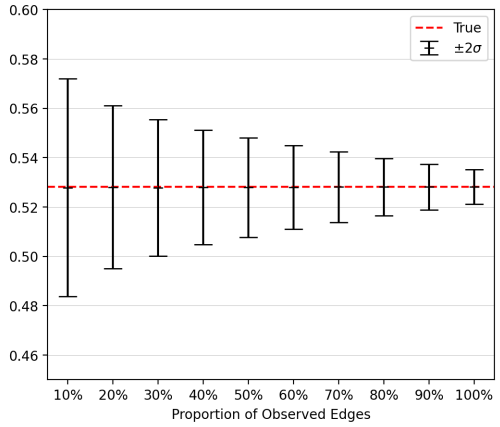
Figure B.2: Simulated Estimates under i.i.d. Assumption
(Linear-in-Sums Model)



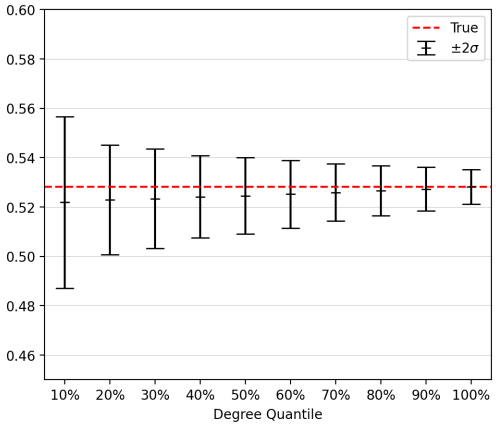
a Estimates for $\hat{\alpha}_1^{obs,sim}$
(Random Missingness)



b Estimates for $\hat{\alpha}_1^{obs,sim}$
(Censored Missingness)



c Estimates for $\hat{\alpha}_2^{obs,sim}$
(Random Missingness)

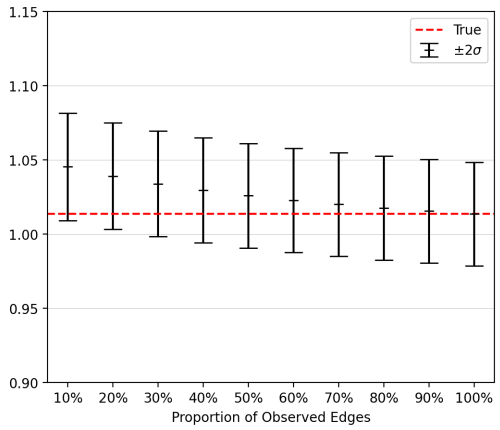


d Estimates for $\hat{\alpha}_2^{obs,sim}$
(Censored Missingness)

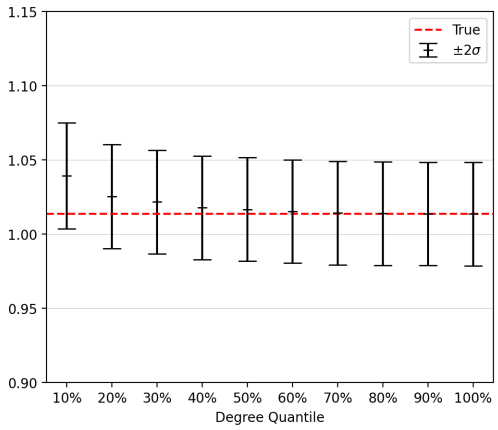
but remain ambiguous for censored missingness. The simulated estimates for $\hat{\alpha}_2^{obs}$ are augmented (larger in magnitude but of the same sign) under random missingness (Panel (c)), while the sign of inconsistency is ambiguous (though augmented in this case) under censored missingness (Panel (d)).[3]

[3]The ambiguity can be tested through a simulation of multiple data generating processes.

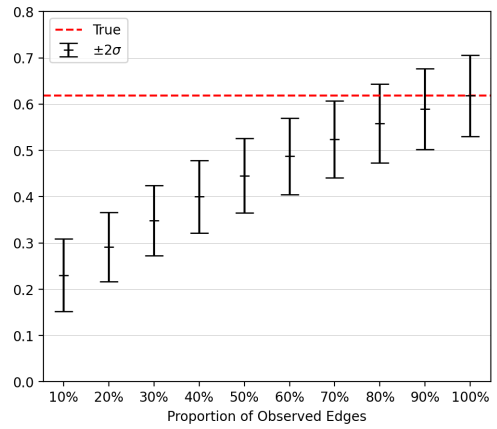
Figure B.3: Simulated Estimates under Weak Homophily Assumption
(Linear-in-Means Model)



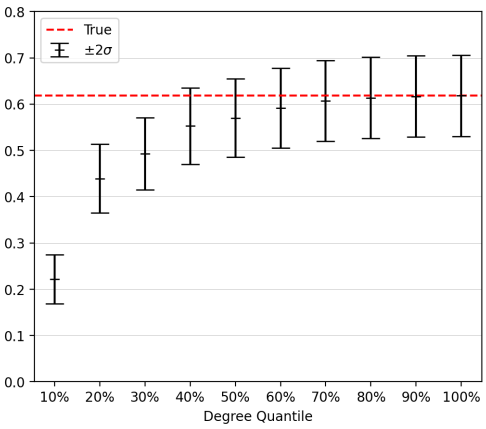
a Estimates for $\hat{\alpha}_1^{obs,sim}$
(Random Missingness)



b Estimates for $\hat{\alpha}_1^{obs,sim}$
(Censored Missingness)



c Estimates for $\hat{\alpha}_2^{obs,sim}$
(Random Missingness)

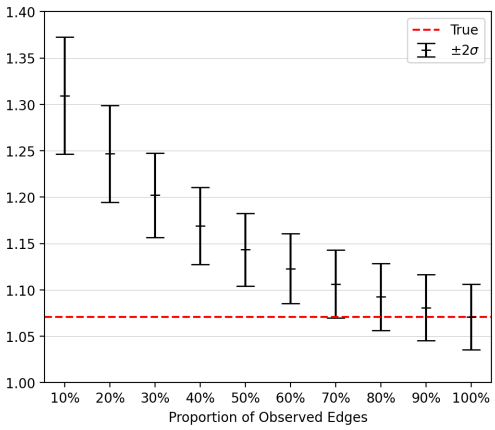


d Estimates for $\hat{\alpha}_2^{obs,sim}$
(Censored Missingness)

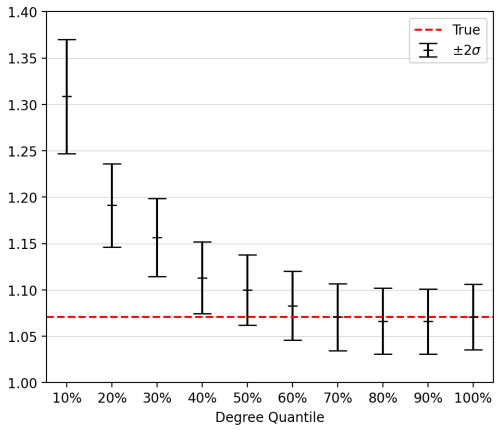
B.2.4 Regressions with Number of Friends

Finally, we show simulated results for regressions that include number of friends as a regressor. As discussed in Subsection 2.5.2, this is a special case of the linear-in-sums case in which the $N \times d$

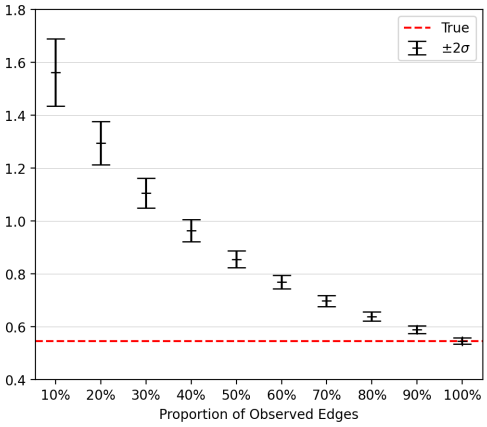
Figure B.4: Simulated Estimates under Weak Homophily Assumption
(Linear-in-Sums Model)



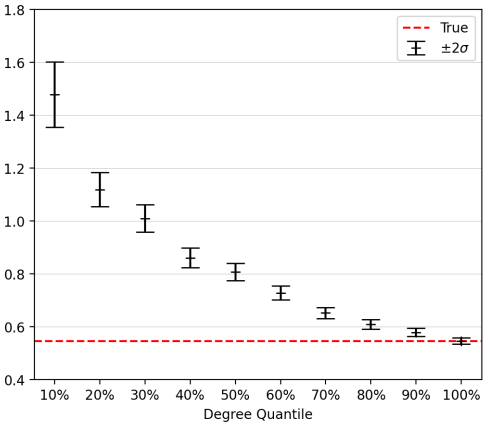
a Estimates for $\hat{\alpha}_1^{obs,sim}$
(Random Missingness)



b Estimates for $\hat{\alpha}_1^{obs,sim}$
(Censored Missingness)



c Estimates for $\hat{\alpha}_2^{obs,sim}$
(Random Missingness)

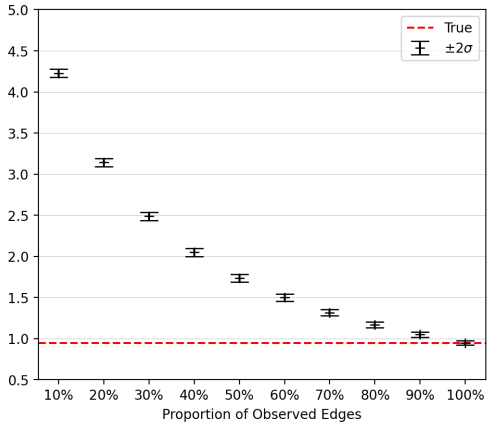


d Estimates for $\hat{\alpha}_2^{obs,sim}$
(Censored Missingness)

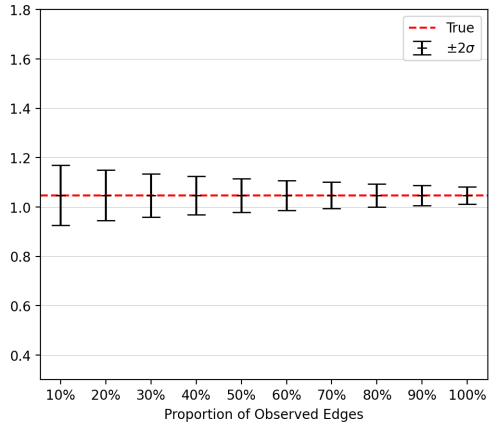
matrix includes a vector of 1's. More specifically, we simulate the mode given in Equation B.144.

$$\mathbf{y} = \iota\alpha_{1,1} + \alpha_{1,2}\mathbf{T} + \mathbf{L}\iota\alpha_{2,1} + \mathbf{LT}\alpha_{2,2} + \varepsilon \tag{B.144}$$

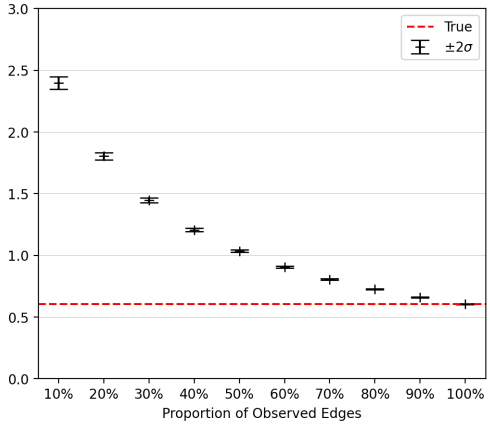
Figure B.5: Simulated Estimates of Regressions with Number of Friends under i.i.d. assumption and Random Missingness
(Linear-in-Sums Model)



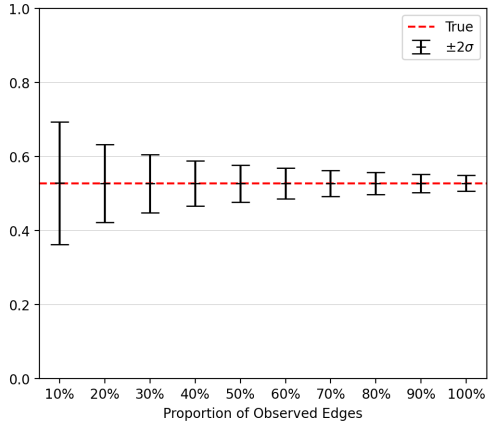
a Estimates for $\hat{\alpha}_{1,1}^{obs,sim}$



b Estimates for $\hat{\alpha}_{1,2}^{obs,sim}$



c Estimates for $\hat{\alpha}_{2,1}^{obs,sim}$



d Estimates for $\hat{\alpha}_{2,2}^{obs,sim}$

In these simulations, $\mathbf{T}$ is a demeaned, randomly-assigned binary treatment status vector. This is equivalent to a particular linear-in-sums model with $d = 2$.[4]

[4]The DGP for the more general case of using $\mathbf{z}$ as in section 2.5.2 would be in form of using a standardized variable of $\tilde{\mathbf{z}} = \mathbf{L}\mathbf{w} + \mathbf{u}$, which not only allows correlation among observations, but also allows the covariates and the number of friends to be correlated. The results shown in Figure B.5 aligns with the findings in Corollary 4 where it shows that the estimates of coefficients on the treatment status $\alpha_{1,2}$ and $\alpha_{2,2}$ are consistent.

We simulate results under random missingness, as that result corresponds to the one given in Corollary 4. These simulations are summarized in Figure B.5. Note that, importantly, the coefficients on own treatment status ($\alpha_{1,2}$) and the number of treated friends ($\alpha_{2,2}$), shown in Panels (b) and (d), are consistent even when many links are missing. The constant term ($\alpha_{1,1}$) and the coefficient on the number of friends ($\alpha_{2,1}$) are both augmented (Panels (a) and (c), respectively), also consistent with Corollary 4.

B.3 Supplemental Tables and Figures

Table B.1: China Insurance Sub-Censored Results ("OR" Network)

Max Number of Links	1	2	3	4	5
Panel A: Means Definition					
$Treat_{is}$ ($\hat{\alpha}_1$)	0.023	0.027	0.034	0.031	0.030
	(0.032)	(0.032)	(0.033)	(0.033)	(0.032)
Mean of Links' $Treat_{js}$ ($\hat{\alpha}_2$)	0.015	0.146*	0.214**	0.279**	0.358***
	(0.049)	(0.075)	(0.098)	(0.112)	(0.124)
R-squared	0.112	0.117	0.126	0.128	0.138
Panel B: Sums Definition					
$Treat_{is}$ ($\hat{\alpha}_1$)	0.024	0.027	0.033	0.030	0.030
	(0.033)	(0.032)	(0.033)	(0.033)	(0.032)
Sum of Links' $Treat_{js}$ ($\hat{\alpha}_2$)	0.003	0.039	0.039	0.044**	0.048**
	(0.042)	(0.028)	(0.024)	(0.021)	(0.019)
R-squared	0.112	0.116	0.124	0.127	0.137
c_1	3.494	1.286	0.571	0.217	0
c_2	6.078	4.393	2.815	1.393	0
c_3	-0.014	-0.005	-0.008	0.002	0

Notes: N = 1,267 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. Standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1.

Table B.2: China Insurance with Randomly Missing Links
("OR" Network)

Percent of Links Observed	25%	50%	75%	100%
Panel A: Means Definition				
$Treat_{is}$ ($\hat{\alpha}_1$)	0.033	0.032	0.033	0.035
	(0.034)	(0.032)	(0.032)	(0.032)
	[0.011]	[0.004]	[0.004]	[0]
Mean of Links' $Treat_{js}$ ($\hat{\alpha}_2$)	0.069	0.116	0.201	0.331
	(0.068)	(0.082)	(0.103)	(0.122)
	[0.058]	[0.060]	[0.057]	[0]
N	1110.411	1273.562	1292.600	1296
Panel B: Sums Definition				
$Treat_{is}$ ($\hat{\alpha}_1$)	0.031	0.032	0.034	0.035
	(0.032)	(0.032)	(0.032)	(0.032)
	[0.003]	[0.003]	[0.004]	[0]
Sum of Links' $Treat_{js}$ ($\hat{\alpha}_2$)	0.038	0.040	0.042	0.046
	(0.037)	(0.026)	(0.022)	(0.019)
	[0.031]	[0.018]	[0.010]	[0]
c_1	2.425	0.965	0.330	0.000
c_2	5.412	3.609	1.805	0.000
c_3	-0.000	-0.000	-0.000	0.000
N	1296	1296	1296	1296

Notes: Based on 1,000 simulations, with links observed with indicated probability. 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. Average standard errors (across simulations) in parentheses, clustered by village. Standard deviation of point estimates (across simulations) in brackets.

Table B.3: 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				
Age	10	19	15.095	1.680
Grade	6	12	9.713	1.584
Female	0	1	0.509	0.500
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

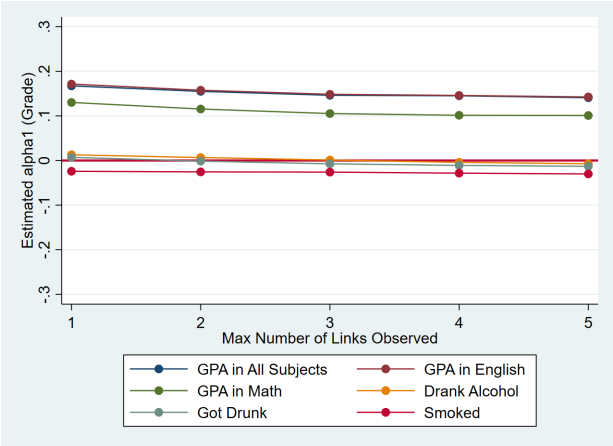
Notes: N = 70,364. Analysis dataset includes only students with friendship nominations and non-missing data for all variables in Panel B.

Table B.4: AddHealth Sub-censored Results
("OUT" Network)

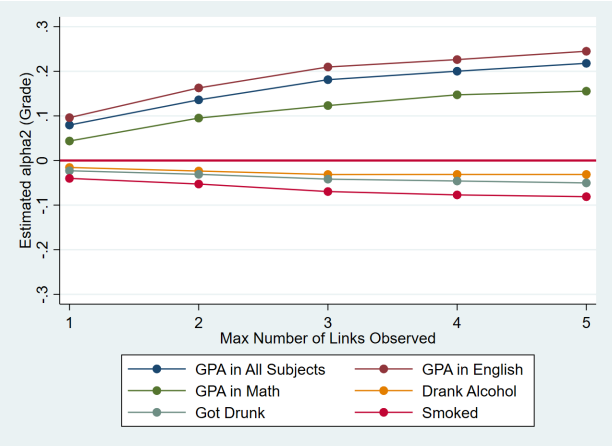
Censoring Rule (k)	1	2	3	4	5
Panel A: Means Specification					
Coefficients on Own Characteristics ($\hat{\alpha}_1$)					
Age	-0.163***	-0.153***	-0.148***	-0.143***	-0.141***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Grade	0.175***	0.170***	0.165***	0.164***	0.163***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
Female	0.148***	0.155***	0.155***	0.154***	0.156***
	(0.012)	(0.012)	(0.013)	(0.013)	(0.013)
Coefficients on Peer Characteristics ($\hat{\alpha}_2$)					
Age	-0.083***	-0.135***	-0.173***	-0.190***	-0.203***
	(0.010)	(0.015)	(0.018)	(0.020)	(0.021)
Grade	0.071***	0.115***	0.151***	0.163***	0.176***
	(0.012)	(0.018)	(0.019)	(0.020)	(0.021)
Female	-0.018	-0.035**	-0.039**	-0.045**	-0.059***
	(0.012)	(0.017)	(0.019)	(0.020)	(0.020)
Panel B: Sums Specification					
Coefficients on Own Characteristics ($\hat{\alpha}_1$)					
Age	-0.164***	-0.156***	-0.152***	-0.150***	-0.148***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Grade	0.184***	0.185***	0.185***	0.183***	0.181***
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Female	0.151***	0.160***	0.159***	0.159***	0.161***
	(0.012)	(0.012)	(0.012)	(0.013)	(0.012)
Coefficients on Peer Characteristics ($\hat{\alpha}_2$)					
Age	-0.045***	-0.035***	-0.029***	-0.024***	-0.022***
	(0.006)	(0.004)	(0.003)	(0.003)	(0.003)
Grade	0.030***	0.023***	0.019***	0.016***	0.015***
	(0.005)	(0.004)	(0.003)	(0.003)	(0.003)
Female	-0.017	-0.024***	-0.018***	-0.016***	-0.017***
	(0.012)	(0.009)	(0.007)	(0.006)	(0.005)

Dependent Variable: GPA in All Subjects. N = 26,465 in all specifications. 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. *** p<0.01, ** p<0.05, * p<0.1.

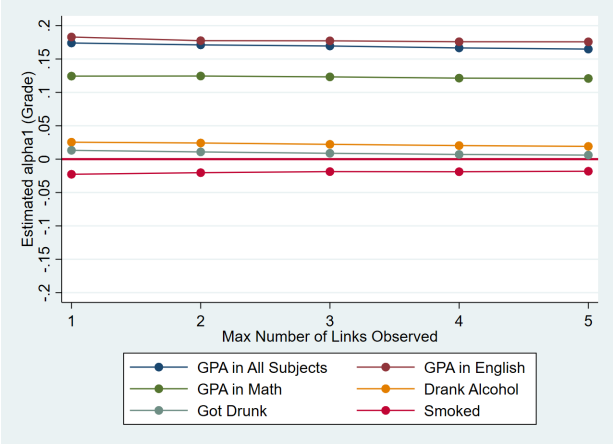
Figure B.6: Sub-censored AddHealth Estimates
("OR" Network, Grade)



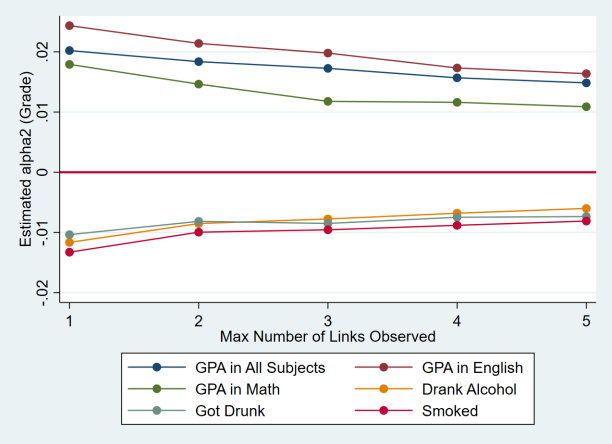
a Coefficients for $Grade_{is}$
(Means Specification)



b Coefficients for $\overline{Grade}_{is}$
(Means Specification)

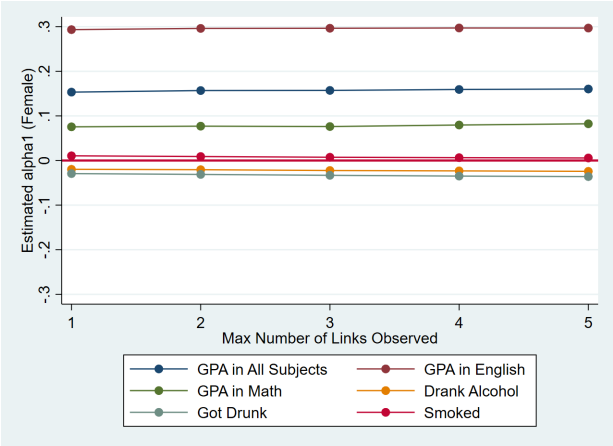


c Coefficients for $Grade_{is}$
(Sums Specification)

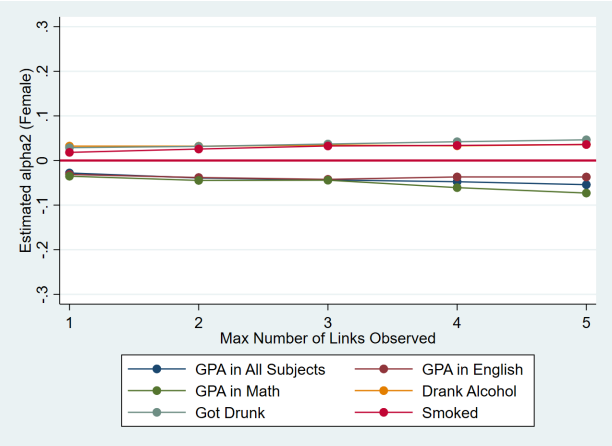


d Coefficients for $\overline{Grade}_{is}$
(Sums Specification)

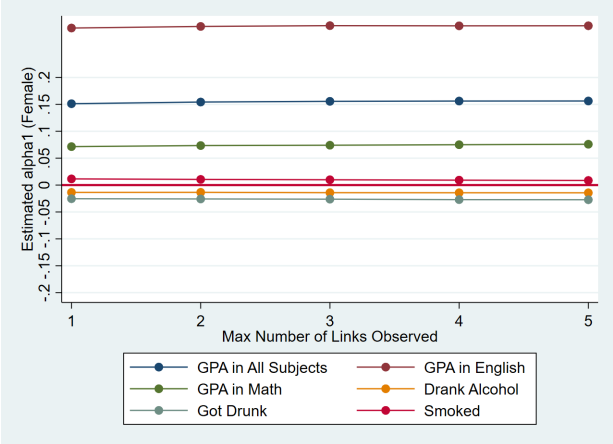
Figure B.7: Sub-censored AddHealth Estimates
("OR" Network, Female)



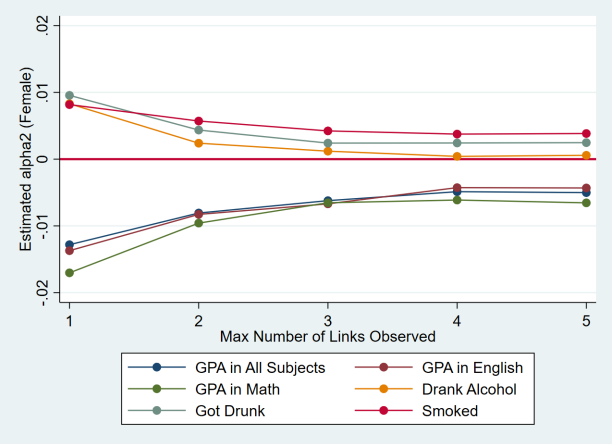
a Coefficients for $Female_{is}$
(Means Specification)



b Coefficients for $\overline{Female}_{is}$
(Means Specification)



c Coefficients for $Female_{is}$
(Sums Specification)



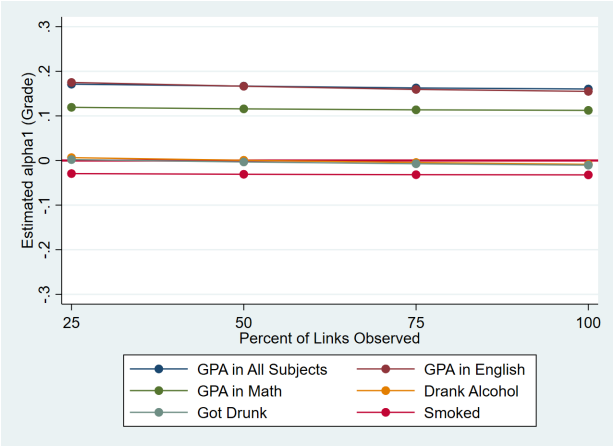
d Coefficients for $\overline{Female}_{is}$
(Sums Specification)

Table B.5: AddHealth Results with Random Missingness
("OUT" Network)

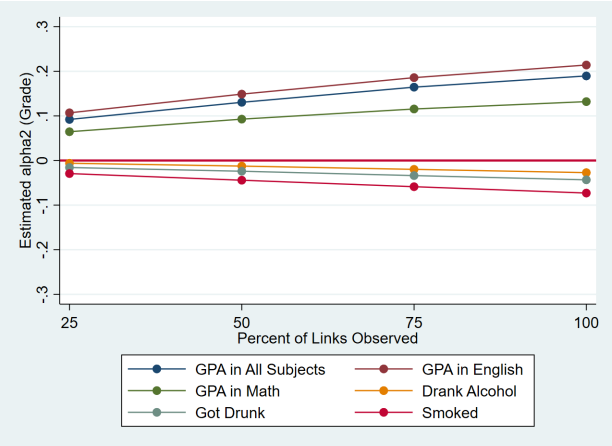
Percent Observed	25%	50%	75%	100%
Panel A: Means Specification				
Coefficients on Own Characteristics ($\hat{\alpha}_1$)				
Age	-0.170	-0.167	-0.163	-0.159
	0.013	0.012	0.012	0.011
Grade	0.184	0.182	0.178	0.174
	0.016	0.015	0.015	0.015
Female	0.150	0.152	0.152	0.151
	0.012	0.011	0.011	0.011
Coefficients on Peer Characteristics ($\hat{\alpha}_2$)				
Age	-0.091	-0.124	-0.156	-0.185
	0.011	0.012	0.014	0.017
Grade	0.076	0.106	0.136	0.164
	0.013	0.013	0.015	0.017
Female	-0.030	-0.034	-0.040	-0.043
	0.012	0.013	0.014	0.016
Observations	24,368.530	32,257.980	35,206.470	36,527.000
Panel B: Sums Specification				
Coefficients on Own Characteristics ($\hat{\alpha}_1$)				
Age	-0.180	-0.176	-0.172	-0.169
	0.012	0.012	0.011	0.011
Grade	0.191	0.189	0.187	0.185
	0.012	0.012	0.012	0.012
Female	0.153	0.155	0.155	0.154
	0.009	0.009	0.009	0.009
Coefficients on Peer Characteristics ($\hat{\alpha}_2$)				
Age	-0.007	-0.006	-0.006	-0.006
	0.002	0.002	0.001	0.001
Grade	0.006	0.004	0.004	0.004
	0.003	0.002	0.002	0.002
Female	-0.006	-0.004	-0.004	-0.002
	0.007	0.006	0.005	0.005
Observations	40,790	40,790	40,790	40,790
c_1	3.000	1.002	0.334	0.000
c_2	5.749	3.834	1.917	0.000
c_3	0.003	0.002	0.001	0.000

Notes: Average point estimates across 1000 simulations. Average standard errors, clustered by school, across simulations in parentheses. Number of observations is average across simulatoins. Within simulations, individuals who have no links for a given draw are omitted in Panel A (but not in Panel B).

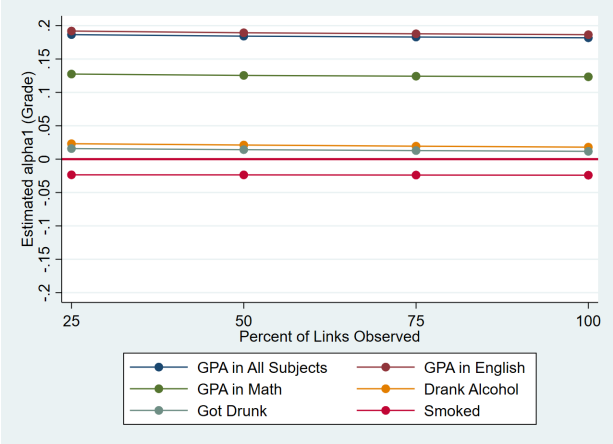
Figure B.8: AddHealth Estimates with Random Missingness
("OR" Network, Grade)



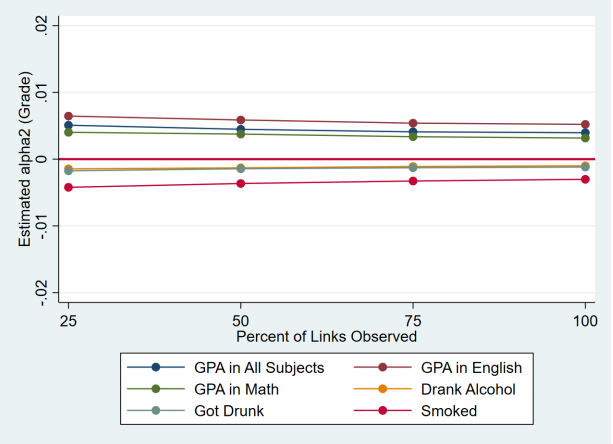
a Coefficients for $Grade_{is}$
(Means Specification)



b Coefficients for $\overline{Grade}_{is}$
(Means Specification)

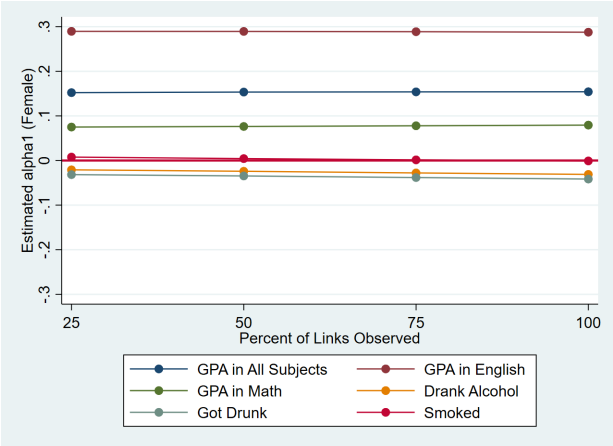


c Coefficients for $Grade_{is}$
(Sums Specification)

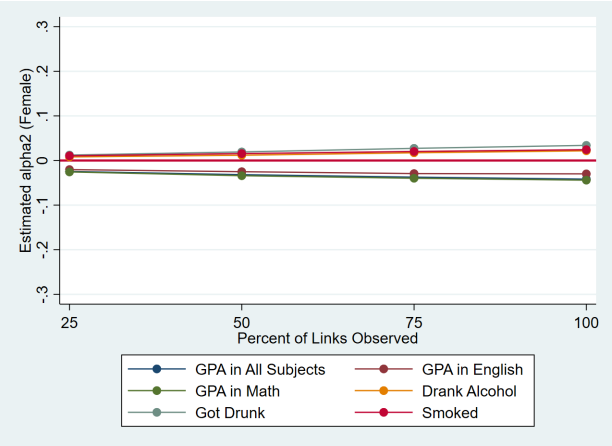


d Coefficients for $\overline{Grade}_{is}$
(Sums Specification)

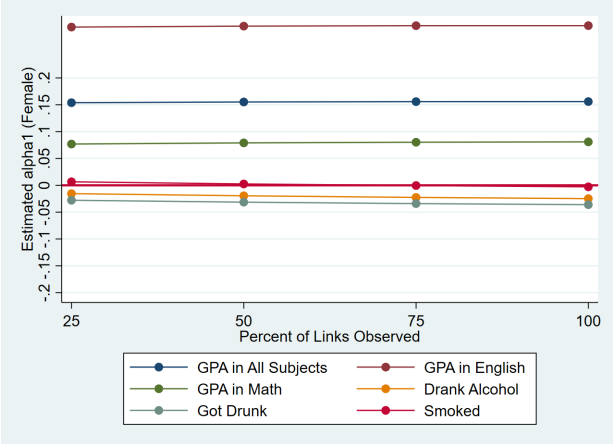
Figure B.9: AddHealth Estimates with Random Missingness
("OR" Network, Female)



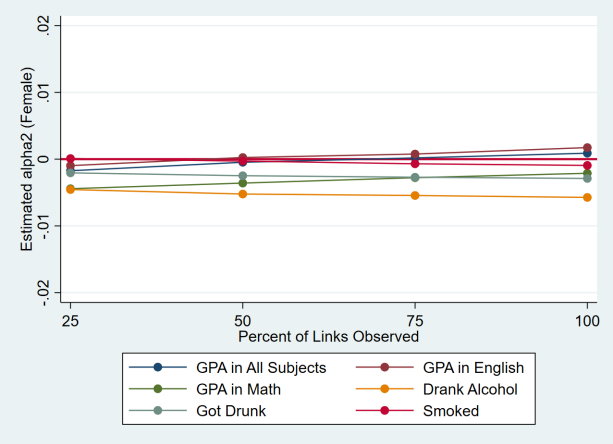
a Coefficients for $Female_{is}$
(Means Specification)



b Coefficients for $\overline{Female_{is}}$
(Means Specification)



c Coefficients for $Female_{is}$
(Sums Specification)



d Coefficients for $\overline{Female_{is}}$
(Sums Specification)

Appendix C

APPENDIX FOR CHAPTER 3

C.1 Proofs

This is a detailed version of section 3.3.

Recall that, for a diffusion process, our objective is to estimate the proportion of agents in each state at each time t. Formally, we estimate $\mathbb{E}[x_j^\tau(t)]$, the mean of a binary valued random variable for each agent j of each state $\tau \in \{S, I, R\}$. For example, in the case of agents in the infectious state ($\tau = I$), we estimate $\mathbb{E}[x_j^I(t)] = i(t)$ where

$$x_j^I(t) = \begin{cases} 1 & \text{if agent } j \text{ is infected} \\ 0 & \text{otherwise.} \end{cases}$$

Assume that $\mathbb{E}\left[x_j^2(t)\right] < \infty$ for all t.

For each epidemic or diffusion process, agents are allocated in to $N_V \geq 1$ disjoint "bins" or "strata" (e.g. defined by degree). For a given set of bins, let q_v be the (known) probability of being in bin v in the population (so, $\sum_v q_v = 1$). Note that all variables in this section are time varying and equivalently applicable for all τ but we drop the notations for simplicity.

C.1.1 Variance Decomposition

For each agent ν_j in stratum v, decompose x_{jv} as follows:

$$x_{jv} = \mu_v + \alpha_v + \epsilon_{jv}$$

Impose the following (innocuous) independence assumptions, which are justified by random sampling within strata:

[1] $\mathbb{E}[\alpha_v] = 0$ for all v

[2] $\mathbb{E}[\epsilon_{jv}] = 0$ for all j, v

[3] $\mathbb{E}[\alpha_v \epsilon_w] = 0$ for all v, w

[4] $\mathbb{E}[\epsilon_{jv}\epsilon_{lw}] = 0$ for all j, l, v, w

Define $\sigma_s^2 = \mathbb{E}\left[\epsilon_{jv}^2\right]$. The independence restrictions above imply that

$$\sigma_s^2 = \mathbb{E}\left[(x_{jv} - \alpha_s)^2\right] + \mathbb{E}\left[\epsilon_{jv}^2\right] = \sigma_{\alpha,v}^2 + \sigma_{\epsilon,v}^2.$$

Now suppose $j, l \neq j$ are sampled. Then there will be correlation in x_{jv} and x_{lv}, which is Suppose that $j, l \neq j$ are sampled. There will be correlation in x_{jv} and x_{lw}, which is dependent upon the underlying parameters of the epidemic/diffusion process. Define

$$\rho_{vw} = \frac{\mathbb{E}\left[(x_{jv} - \mu_v)(x_{lw} - \mu_w)\right]}{\sigma_v \sigma_w} \qquad \forall j \neq l$$

clearly, for all v, w, $\rho_{vw}\sigma_v\sigma_w = \mathbb{E}[\alpha_v \alpha_w]$. Additionally, when $v = w$,

$$\rho_{vv} = \frac{\sigma_{\alpha,v}^2}{\sigma_{\alpha,v}^2 + \sigma_{\epsilon,v}^2}$$

Optimal Stratified Sampling

For a fixed sample size N, let N_v be the number of agents in each stratum v where $\sum_v N_v = N$. So, the estimator of μ is

$$\hat{\mu}(\mathbf{N}) = \sum_v q_v \left(\frac{1}{N_v} \sum_{j=1}^{N_v} x_{jv} \right).$$

where $\mathbf{N} = (N_1, ..., N_V)$. Since $\mathbb{E}[x_{jv}] = \mu_v$ for all v, $\hat{\mu}(\mathbf{N})$ is always unbiased. The variance is given by

$$\mathbb{E}\left[(\hat{\mu}(\mathbf{N}) - \mu)^2\right] = \mathbb{E}\left[\left(\sum_v q_v \left(\frac{1}{N_v}\sum_{j=1}^{N_v} x_{jv}\right) - \mu\right)^2\right]$$

which can be rewritten as

$$\begin{aligned}
\mathbb{E}\left[(\hat{\mu}(\mathbf{N}) - \mu)^2\right] &= \sum_v \sum_w \frac{q_v q_w}{N_v N_w} \sum_j \sum_l \mathbb{E}\left[(x_{jv} - \mu_v)(x_{lw} - \mu_w)\right] \\
&= \sum_v \frac{q_v^2}{N_v^2} \sum_j \sum_l \mathbb{E}\left[(x_{jv} - \mu_v)(x_{lw} - \mu_w)\right] \\
&\quad + \sum_v \sum_{w \neq v} \frac{q_v q_w}{N_v N_w} \sum_j \sum_l \mathbb{E}\left[(x_{jv} - \mu_v)(x_{lw} - \mu_w)\right] \\
&= \sum_v \frac{q_v^2}{N_v^2} \left[\sum_j \mathbb{E}\left[(x_{jv} - \mu_v)^2\right] + \sum_j \sum_{j \neq l} \mathbb{E}\left[(x_{jv} - \mu_v)(x_{lv} - \mu_v)\right]\right] \\
&\quad + \sum_v \sum_{w \neq v} \frac{q_v q_w}{N_v N_w} \sum_j \sum_l \mathbb{E}\left[(x_{jv} - \mu_v)(x_{lw} - \mu_w)\right] \\
&= \sum_v \frac{q_v^2}{N_v^2} \left[\sigma_v^2 \left(1 + (N_v - 1)\sigma_s^2 \rho_{vv}\right)\right] + \sum_v \sum_{w \neq v} q_v q_w \rho_{vw} \sigma_v \sigma_w \\
&= \sum_v \frac{q_v^2 \sigma_v^2 (1 - \rho_{vv})}{N_v} + \sum_v \sum_w q_v q_w \rho_{vw} \sigma_v \sigma_w
\end{aligned}$$

We can minimize this subject to the following constraints:

$$\sum_v N_v = N$$

$$N_v \geq 1 \quad \forall v$$

which yields the following FOCs for each v:

$$\frac{q_v^2 \sigma_v^2 (1 - \rho_{vv})}{N_v^2} = \lambda - \mu_v$$

where $\lambda, \mu_v \geq 0$. The solution to this is N_v^* for all v as follows:

$$N_v^* = \frac{q_v \sigma_v \sqrt{1-\rho_{vv}}}{\sum_w q_w \sigma_w \sqrt{1-\rho_{ww}} N}$$

Since $1 - \rho_{vv} = \frac{\sigma_{\epsilon,v}^2}{\sigma_v^2}$, this can be restated as

$$N_v^* = \frac{q_v \sigma_{\epsilon,v}}{\sum_w q_w \sigma_{\epsilon,w}} N.$$

where $N_v^* \geq 1$ for all v for $\hat{\mu}\left(\mathbf{N}^*\right)$ to be unbiased. Therefore, the optimal allocation depends only on the within-strata variance. Further, the variance at $\mathbf{N}^*$ is given by:

$$\mathbb{V}\left[\hat{\mu}\left(\mathbf{N}^*\right)\right] = \frac{1}{N}\left(\sum_v q_v \sigma_{\epsilon,v}\right)^2 + \sum_v \sum_w q_v q_w \rho_{vw} \sigma_v \sigma_w$$

Simple Random Sampling

First, we compare simple random sampling, where N are drawn at random from the population. Define this estimator as $\hat{\mu}_{SRS}$. So,

$$\begin{aligned}
\mathbb{V}\left[\hat{\mu}_{SRS}\right] &= \mathbf{E}\left[\left(\hat{\mu}_{SRS} - \mu\right)^2\right] \\
&= \mathbf{E}\left[\left(\frac{1}{N}\sum_j x_{jv} - \mu\right)\left(\frac{1}{N}\sum_l x_{lw} - \mu\right)\right] \\
&= \frac{1}{N^2}\sum_j \mathbb{E}\left[\left(x_{jv} - \mu\right)^2\right] + \frac{1}{N^2}\sum_j \sum_{l \neq j} \mathbb{E}\left[\left(x_{jv} - \mu\right)\left(x_{lw} - \mu\right)\right] \\
&= \frac{1}{N}\left[\sum_v q_v \sigma_v^2 + (N-1)\sum_v \sum_w q_v q_w \rho_{vw} \sigma_v \sigma_w\right] \\
&= \frac{1}{N}\sum_v q_v\left(\sigma_{\alpha,v}^2 + \sigma_{\epsilon,v}^2\right) + \frac{N-1}{N}\sum_v \sum_w q_v q_w \rho_{vw} \sigma_v \sigma_w
\end{aligned}$$

Stratified Random Sampling

Next, define unbiased stratified random sampling, where $q_v = \frac{N_v}{N}$ for all v. That is, the sample is divided up according to the proportions of agents in each stratum in the population. Let $\mathbf{q}$ be the vector of frequencies in each stratum in the population. Therefore $\mathbf{N} = N\mathbf{q}$. So,

$$\mathbb{V}\left[\hat{\mu}\left(N\mathbf{q}\right)\right] = \mathbb{E}\left[\left(\hat{\mu}\left(N\mathbf{q}\right) - \mu\right)^2\right] = \frac{1}{N}\sum_v q_v \sigma_{\epsilon,v}^2 + \sum_v \sum_w q_v q_w \rho_{vw} \sigma_v \sigma_w$$

Comparisons

The gains in variance moving from SRS to unbiased stratified random sampling are given as

$$\mathbb{V}\left[\hat{\mu}_{SRS}\right] - \mathbb{V}\left[\hat{\mu}\left(N\mathbf{q}\right)\right] = \frac{1}{N}\left(\sum_v q_v \sigma_{\alpha,v}^2 - \sum_v \sum_w q_v q_w \rho_{vw} \sigma_v \sigma_w\right)$$

This can be restated as

$$\mathbb{V}\left[\hat{\mu}_{SRS}\right] - \mathbb{V}\left[\hat{\mu}\left(N\mathbf{q}\right)\right] = \frac{1}{N}\left(\mathbb{E}_v\left[\sigma_{\alpha,v}^2\right] - \mathbb{E}_{v,w}\left[\alpha_v \alpha_w\right]\right)$$

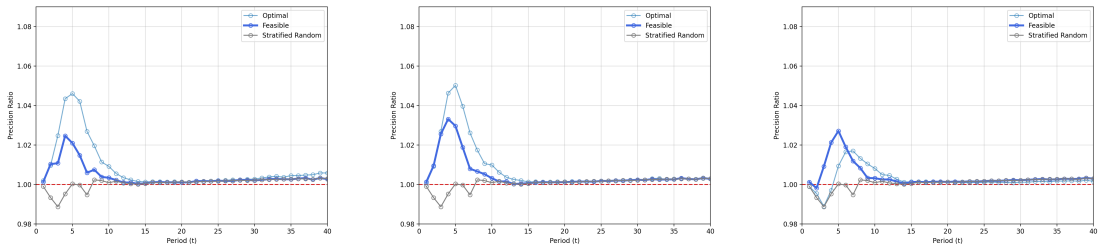
which must always be (weakly) positive due to the Cauchy Schwartz Inequality. Next,

$$\mathbb{V}\left[\hat{\mu}\left(N\mathbf{q}\right)\right] - \mathbb{V}\left[\hat{\mu}\left(\mathbf{N}^*\right)\right] = \frac{1}{N}\left(\mathbb{E}_v\left[\sigma_{\epsilon,v}^2\right] - \mathbb{E}_v\left[\sigma_{\epsilon,v}\right]^2\right) = \frac{1}{N}\mathbb{V}_v\left[\sigma_{\epsilon,v}\right]$$

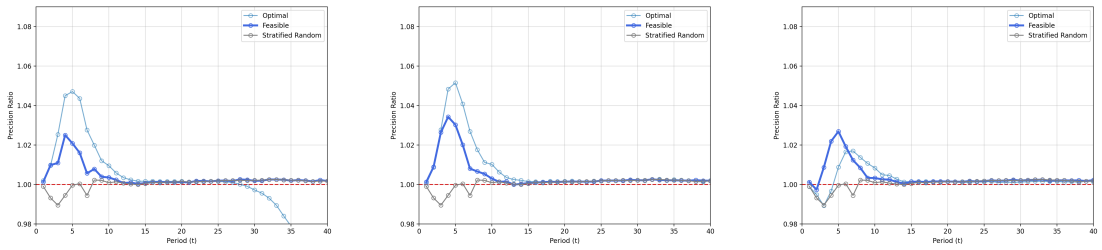
That is, the gain for the optimal stratified sampling is determined by the variance in the unobserved variance $\mathbb{V}_v\left[\sigma_{\epsilon,v}\right]$. In the corner case where this is zero, there is no gain.

C.2 Supplementary Figures

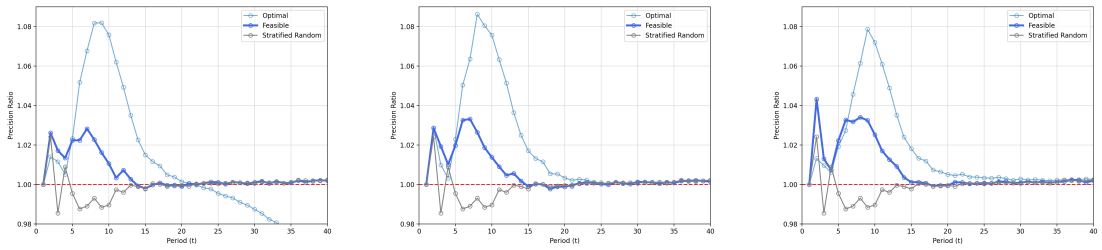
a $S\,(\tau = S)$ b $I\,(\tau = S)$ c $R\,(\tau = S)$



d $S\,(\tau = I)$ e $I\,(\tau = I)$ f $R\,(\tau = I)$



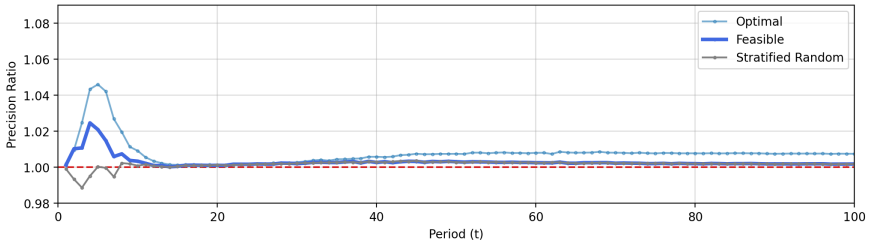
g $S\,(\tau = R)$ h $I\,(\tau = R)$ i $R\,(\tau = R)$



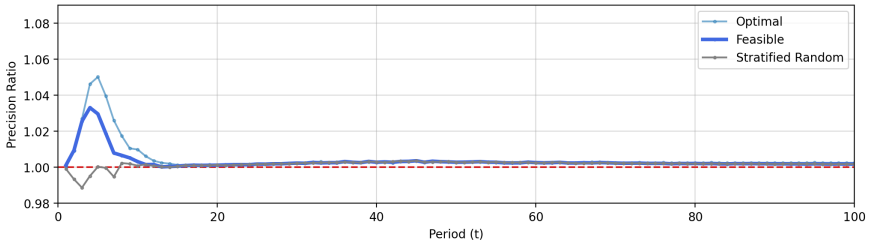
Note: Each line represents the precision ratio of each sampling method to random sampling.

Figure C.1: SIR Model Sampling Performance Results for each state of early periods and all $\tau \in S, I, R$

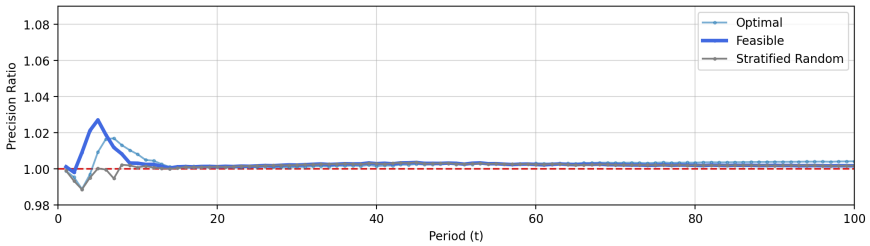
a SIR – S ($\tau = S$)



b SIR – S ($\tau = I$)



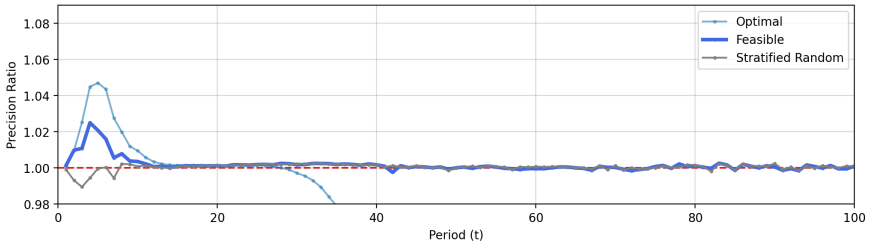
c SIR – S ($\tau = R$)



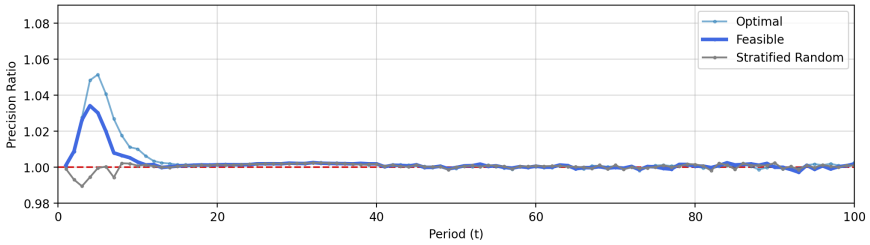
Note: Each line represents the precision ratio of each sampling method to random sampling.

Figure C.2: SIR Model Sampling Performance Results for each state of all periods and $\tau = S$

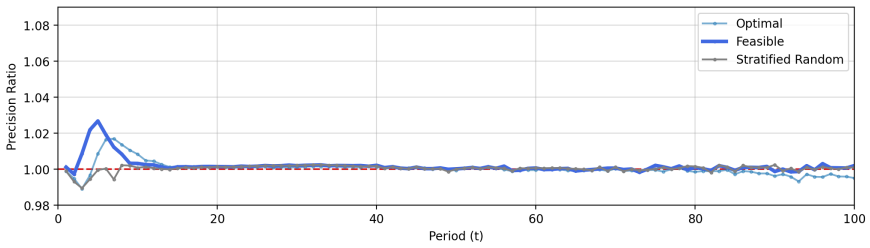
a SIR – I ($\tau = S$)



b SIR – I ($\tau = I$)



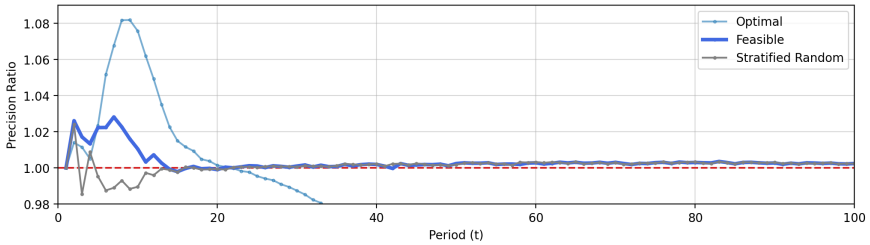
c SIR – I ($\tau = R$)



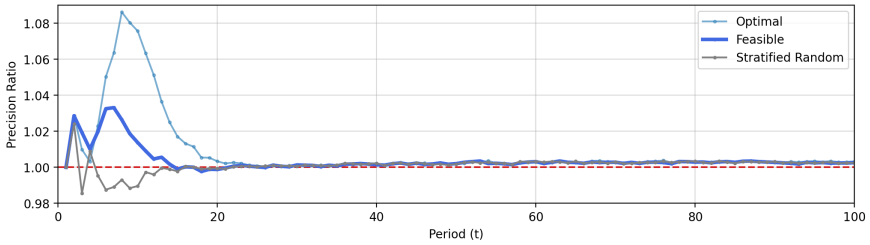
Note: Each line represents the precision ratio of each sampling method to random sampling.

Figure C.3: SIR Model Sampling Performance Results for each state of all periods and $\tau = I$

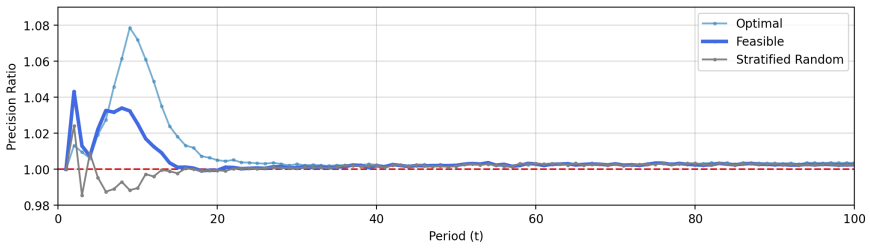
a SIR – R ($\tau = S$)



b SIR – R ($\tau = R$)



c SIR – R ($\tau = R$)



Note: Each line represents the precision ratio of each sampling method to random sampling.

Figure C.4: SIR Model Sampling Performance Results for each state of all periods and $\tau = I$