Improving control over unobservables with network data*

Vincent Starck†

November 20, 2022

Latest version

Abstract

Unobserved variables often threaten the causal interpretation of empirical estimates. An opportunity to alleviate this concern lies in network datasets, which provide a rich source of information about individual characteristics insofar as they influence network formation. This paper develops the idea of controlling for unobserved confounders by leveraging network structures that exhibit homophily, a frequently observed tendency to associate with similar people. This is formally accomplished under two main frameworks. First, I introduce a concept of *strong homophily*, according to which individuals' selectivity is at scale with the size of the potential connection pool, and I show that an estimator that considers neighbors as a comparison group is consistent for the Conditional Average Treatment Effect (CATE). I then consider a setting without *strong homophily* and show how selecting connected individuals whose observed characteristics made such a connection less likely delivers an estimator with similar properties. Overall, the method allows nonparametric treatment effect inference for both CATE and Average Treatment Effect (ATE) under a version of unconfoundedness that conditions on unobservables, which is often more credible than selection on observables alone. In an application, I recover an estimate of the effect of parental involvement on students' test scores that is greater than that of OLS, due to the estimator's ability to account for unobserved ability and effort.

Keywords: Selection on unobservables, Networks, Homophily, Treatment effect.

*I am very grateful to Susanne M. Schennach for her helpful comments and continuous support throughout this project. I also thank Toru Kitagawa, Soonwoo Kwon, and Jonathan Roth for their numerous suggestions and valuable advice. This research has also benefited from seminar participants at Brown.

†Brown University, Department of Economics, 64 Waterman Street, Providence, RI, USA. vincent_starck@brown.edu

1 Introduction

Estimating the effect of a treatment is a frequent goal in economics and social sciences. A common challenge is the presence of unobserved confounders that threaten the validity of the unconfoundedness assumption, which is typically necessary to perform inference with standard methods. Tools that strengthen control over variables that affect outcomes but are hard to measure, such as ability, culture, work ethic, tastes, *etc.*, are thus particularly valuable.

Recently, networks and datasets with a spatial structure have become increasingly available to researchers, providing new avenues for research. Homophily or assortative matching is a ubiquitous feature of empirical networks: individuals tend to associate with similar individuals (Clark and Ayers, 1992; McPherson, Smith-Lovin and Cook, 2001; Jackson, 2010; Currarini, Jackson and Pin, 2009; Moody, 2001; Boucher and Mourifié, 2017; Dzemski, 2019).

Homophily creates opportunities to create unobservable-adjusted comparison groups. For instance, if we are interested in the effect of parental involvement on student test scores, we may be concerned about unobserved confounding from differences in student ability. However, if students of similar ability are more likely to be friends with each other (Clark and Ayers, 1992; Burgess et al., 2011; Boutwell, Meldrum and Petkovsek, 2017), the omitted variable bias can be reduced by comparing connected students.

The paper develops the idea that homophilic networks can be exploited to derive a consistent estimator of treatment effects in the presence of unobserved confounders. This is formally done under two main frameworks: either the network is strongly homophilic – homophilic behavior is the core feature of network formation and is pronounced relative to sample size – or the network at least features homophily in unobservables.

In the former case, I let the link formation process vary with the size of the network; people become pickier to limit the number of connections or improve their average quality as the network expands. As people are able to form increasingly better matches with a larger pool of potential neighbors, they become more selective because of decreasing benefits per additional match, preference for quality of matches, or limited resources to devote to additional connections.

This accomplishes two things. First, the approach provides an asymptotic approximation that preserves some degree of sparsity and that does not render the mechanism of network formation negligible in the limit; selectivity is at scale with the size of the connection pool. As such, it contributes to the literature on network formation. Second, and most importantly, it establishes consistency and asymptotic normality for the proposed intuitive estimators that use (possibly high-order) connections or connections in common as comparison groups.

In an alternative framework where homophily is not strong, I impose homophily only in the unobserved variables. This allows an arbitrary functional form for the observed covariates in link formation and, possibly, a link formation function independent of the size of the network. In this scenario, comparison groups can be derived by comparing people who are different on observables but connect nevertheless. This hinges on the following intuition: if there is no observed rationale for two people being friends, the reason for their friendship likely lies in the unobserved world. If two people are connected despite their observables indicating such a link was unlikely, they are more likely to be close in terms of unobservables. By suitably manipulating the allowed discrepancy in observables relative to sample size, one can recover consistent estimators.

I provide results that allow for the estimation of the Conditional Average Treatment Effect (CATE), which provides a way to describe the heterogeneity of the treatment effect for sampled individuals. The conditional average effect may be the end goal of the analysis (when a specific unit is targeted for treatment or policy) or may be a prelude to aggregation to the Average Treatment Effect (ATE).

I define a general form of CATE estimator as a function of a group of counterfactual observations to be determined, then propose different choices to deal with different empirical issues. In all cases, estimators isolate increasingly better counterfactuals as to recover the CATE asymptotically. I show that the proposed estimators of the (C)ATE are asymptotically normal, enabling statistical inference.

Although results pertain to nonparametric estimators, the intuition remains valid in parametric specifications and similar results are typically achievable under similar or weaker conditions. Propensity score analysis – and then possibly doubly robust estimators – can also be analyzed analogously.

Finally, I demonstrate the feasibility and effectiveness of the method through both simulations and an empirical application. In the application, I obtain an estimate of the effect of parental involvement on students' test scores that suggests a greater impact than OLS does, due to the estimator's ability to account for unobserved ability and effort.

1.1 Related literature

The paper is at the intersection of the literature on networks (Jackson, 2010; Graham, 2015; De Paula, 2017), in particular those featuring homophilic network formation (Boucher, 2015; Graham, 2016; Demirer, 2019), and estimation of treatment effects (Imbens, 2004; Imbens and Wooldridge, 2009; Imbens and Rubin, 2015), both of which considerably grew in size over the last decades.

A closely related paper is Auerbach (2019, 2022) who considers a partially linear outcome regression where the nonlinear term depends on an unobserved variable.

Using information from a network whose formation hinges on the unobserved variable, he was able to recover consistent estimates of regression coefficients under general assumptions. See also Johnsson and Moon (2021) who use a similar framework and analyze peer effects and Demirer (2019) who provides partial identification results in linear models under homophilic behavior and proposes new definitions of homophily.

The present paper considers a more general outcome equation in a potential outcome framework, at the expense of generality in the network formation process. Specifically, I consider a nonparametric treatment effect setup, but I impose more structure on network formation, especially homophily in the unobservables. The nonparametric approach allows us to deal with the common concerns of treatment effect heterogeneity and nonlinearities. In addition, the method circumvents the need to define and estimate equivalent classes and allows the use of higher-order neighbors or friends in common through triangular inequality relationships. This allows for intuitive estimators that are easy to implement.

2 Improving control over unobservables using network data

2.1 Notation and assumptions

The sample is a cross-section of n individuals. The treatment status of individual i, $T_i \in \{0, 1\}$, and the corresponding outcome, $Y_i(T_i)$ in potential outcome notation (Rubin, 1974), are observed. As the notation for the outcome suggests, the Stable Unit Treatment Value Assumption (SUTVA) is made[1].

The covariates, $X = (X^o, X^u) \in \mathcal{X}^o \times \mathcal{X}^u = \mathcal{X} \subset \Omega$, are divided into observed variables, X^o, and unobserved variables, X^u. Ω is a vector space with norm $\|\cdot\|$ (with some abuse of notation, this will be used to represent the norm on $\mathcal{X}^o$ or $\mathcal{X}^u$), typically Euclidean, and $\mathcal{X}$ is an open subset of Ω. I focus on continuously distributed covariates X, though discrete variables can be accommodated – typically under weaker conditions since concerns such as asymptotic bias disappear. The

[1]The possible existence of peer effects or interactions among individuals requires careful modelling, possibly starting with a modification of what a potential outcome is (*e.g.*, Forastiere, Airoldi and Mealli (2020)). In particular, the problem of interference has recently received attention and a growing literature addresses (versions of) the issue by refining the propensity score (Jackson, Lin and Yu, 2020; Sánchez-Becerra, 2021; Aronow and Samii, 2017; Arpino, Benedictis and Mattei, 2017; Liu et al., 2019; Sofrygin and van der Laan, 2016; Forastiere, Airoldi and Mealli, 2020). Although homophilic restrictions may still carry identifying power in such settings, deriving estimators and establishing their properties in frameworks with interference is beyond the scope of this paper.

density of a random variable is denoted by f with the corresponding subscript. $B_r(x)$ denotes a ball of radius r centered at x.

Draws of (Y_i, T_i, X_i) are i.i.d. and realizations of a random variable are denoted by the corresponding lower-case letter. C represents a generic constant satisfying $0 < C < \infty$.

A spatial structure or a network is given through a (binary) weighting/link matrix W, of size $(n \times n)$. The neighborhood $\mathcal{N}(i)$ refers to the links, friends, or connections of individual i, *i.e* $\mathcal{N}(i) \stackrel{\text{def}}{=} \{j \in \{1, \ldots, n\} | W_{ij} = 1\}$, and $\mathcal{N}_t(i)$ denotes friends with a specific treatment status t, *i.e* $\mathcal{N}_t(i) \stackrel{\text{def}}{=} \{j \in \{1, \ldots, n\} | W_{ij} = 1, T_j = t\}$. Higher-order friends, say of order m, are denoted by $\mathcal{N}_t^m(i) \stackrel{\text{def}}{=} \{j \in \{1, \ldots, n\} | \exists j_0 = i, \ldots, j_m = j,\ W_{j_k j_{k+1}} = 1, W_{j_k j_l} = 0\ \forall k = 0, \ldots, m, l \neq k+1, T_j = t\}$, and friends in common are given by $\mathcal{N}_t(i; j) \stackrel{\text{def}}{=} \mathcal{N}_t(i) \cap \mathcal{N}_t(j)$. A c superscript is used to denote the complement of a set.

The goal is to conduct inference about the Conditional Average Treatment Effect (CATE) function: $\text{CATE}(x) \stackrel{\text{def}}{=} \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x]$ at some x of positive density (more precisely, it is assumed that there exists $\overline{f}$, $\underline{f}$, and $\underline{r}$ such that $\overline{f} > f > \underline{f} > 0$ on some ball $B_{\underline{r}}(x)$. With $f_X(x) > 0$ on the open set $\mathcal{X}$, continuity of the density is a sufficient condition.). The usual statement about omitting 'almost surely' qualifiers, in particular pertaining to conditional expectations, applies.

The following core assumptions are maintained throughout the paper:

Assumption 2.1 (Core)**.** a) Unconfoundedness: $(Y_i(1), Y_i(0)) \perp T_i | X_i$
b) Overlap: $0 < C < \mathbb{P}[T_i = 1 | X_i] < 1 - C < 1$

These two assumptions are ubiquitous in the treatment effect literature, though this version of unconfoundedness conditions on X instead of X^o. It is thus only assumed that treatment is independent of potential outcomes when conditioned on individual characteristics, including unobserved ones. Since covariates that may influence selection into treatment such as ability, work ethic, or personal preferences are typically unobserved, this is often a valuable relaxation: selection on some unobservables is allowed.

2.2 Network formation

Let i, j be two individuals and $i \neq j$. I focus on link-formation models of the type

$$W_{ij} = \mathbb{1}_{w_n(h(X_i^o; X_j^o) + \|X_i^u - X_j^u\|) \leq \eta_{ij}} \tag{1}$$

where $w_n : \mathbb{R}^+ \to [0; 1]$ is an increasing function that satisfies $\lim_{x \to \infty} w_n(x) = 1$ and may vary with network size (typically, w_n would increase with n to accommodate some sparsity in the network, e.g., $w_n(x) = \min\{nx, 1\}$ or $1 - e^{-nx^2}$). The function

h is arbitrary but known and $\eta_{ij} = \eta_{ji}$ are independent uniform[2] shocks, drawn independently of $(X_i, X_j, T_i, T_j, Y_i, Y_j)$.

Such dyadic network formation processes are common in the literature, *e.g.*, Graham (2014); Auerbach (2019); Johnsson and Moon (2021). Compared to the most general forms specifications (for instance, Auerbach (2019) uses $\mathbb{1}_{w(X_i,X_j)\leq\eta_{ij}}$ and only makes a weak continuity assumption on w), model (1) adds some separability and homophily in unobservables. Although the homophilic structure is an additional assumption, it may often be plausible given the empirical prevalence of homophily. Another departure from the literature is the dependence of w on sample size. I will provide an explicit model of such dependence, which is useful to describe sorting behavior that hinges on the number of potential connections and to limit the asymptotic density of the network.

The function h may feature homophily as well[3], in which case $h(X_i^o; X_j^o) = \|X_i^o - X_j^o\|$ as in Subsection 2.3. In some applications, however, it may be of interest to allow for another functional form and thus to leave h flexible, as in Subsection 2.4. For instance, some work relationships may warrant complementarity in background education, in which case there is a non-(possibly anti-) homophilic selection in a covariate.

The model can be given the usual interpretation of 'link creation under a mutual positive utility of forming a link' ($w - \eta$ then reflecting utility), see, *e.g.*, Jackson (2010); Goldsmith-Pinkham and Imbens (2013), or rationalize the idea that people with similar characteristics are more likely to meet and thus to form a connection. Nevertheless, since the network is primarily seen as information to draw from, this rationale may not be necessary. For instance, if individuals end up developing similar characteristics after randomly forming connections, a researcher that observes the network after covariates have evolved could use the present methods. In other words, (1) need not be the structural equation for network formation but should approximate the relationship between links and covariates at the time of observation.

I first explore the case of *strong homophily*, *i.e.*, homophilic behavior is the core mechanism of network formation and the selectivity of the individuals is tied to the size of their potential matching pool.

The *strong homophily* case provides an asymptotic theory when homophilic behavior is pronounced relative to network size and formalizes the intuition that connections among individuals can be used to form comparison groups. It also shows

[2]Since one can apply an inverse cumulative distribution function on both sides to induce any distribution, the uniform assumption is made without loss of generality.

[3]In this case in particular, it may make sense to consider variables whose variance has been normalized to put them on the same scale. Nevertheless, the results hold if the norms weight each dimension differently as to reflect stronger selection in some covariates.

the identifying power of homophilic restrictions under simpler conditions than the results of the later sections and applies even if only the network structure and the outcome were observed.

An alternative framework is later provided in Subsection 2.4, possibly letting the link formation be independent of sample size. Both cases are compatible with some degree of sparsity in the network. In the first case, increased selectivity as n rises naturally leads to a diminishing connections-to-sample-size ratio. In both cases, it is possible to include variables such as a physical distance that decreases the likelihood of a connection between the newly added observations and previous ones, in a way to be tailored to the unfolding geography of the application.

2.3 Strong homophily

2.3.1 The strong homophily framework

Suppose people associate based on homophily with $h(X_i^o; X_j^o) = \|X_i^o - X_j^o\|$ and, as the network expands, they become pickier as to limit the number of connections or improve their quality. As people are able to form increasingly better matches with a larger pool of potential neighbors, they may become more selective because of decreasing benefits per additional match, preference for quality of matches, or limited resources to devote to additional connections.

To reflect this behavior, the sequence of functions w_n must satisfy two conditions. First, the sequence must be increasing in order to decrease the probability of forming connections as n rises. Homophily further suggests that, in addition, people penalize dissimilar individuals increasingly more harshly so that the average match quality (in terms of homophilic preferences) increases.

Functions of the form $w_n(x) \leq 1 - g(s_n x)$ are consistent with such behavior – homophily becomes more prevalent as n rises – irrespective of the exact form of w_n (or g). Although the weak equality suffices for some results, it will be convenient to replace it by an equality and to impose some regularity conditions for asymptotic results. I thus adopt the following definitions:

Definition 2.1 (Strong Homophily)**.** a) The network formation is strongly homophilic if (1) holds with $h(X_i^o; X_j^o) = \|X_i^o - X_j^o\|$, $w_n(x) \leq 1 - g(s_n x)$, where $g : [0; \infty[\to [0; 1]$ is decreasing and $\lim_{n\to\infty} s_n = \infty$.
b) The network formation is regularly strongly homophilic if (1) holds with $h(X_i^o; X_j^o) = \|X_i^o - X_j^o\|$, $w_n(x) = 1 - g(s_n x)$, where $g : [0; \infty[\to [0; 1]$ is a decreasing function such that $0 < \int_\Omega g(\|y\|)\, dy < \infty$, and $\lim_{n\to\infty} n s_n^{-d} = \infty$.

The definition is new and encodes the fact that many networks exhibit some sparsity, suggesting the drag on connection probability brought by $w_n \uparrow 1$ pointwise, along with intuition of homophilic matching. Part a) formalizes the idea of

strong homophily; part b) provides supplementary conditions under which asymptotic results are available.

Increasing homophily, if not taken literally, formalizes the notion that homophilic behavior is pronounced relative to sample size in the spirit of a drifting sequence; the degree to which individuals are selective is in a sense preserved as we proceed to an asymptotic approximation. These asymptotics thus provide a good approximation to the finite sample properties of the data in settings where (i) sample averages are taken over enough observations for law-of-large-numbers approximations to be meaningful and (ii) the degree of sorting that individuals applied when selecting matches was sizable compared to the pool of potential connection. In particular, strong homophily asymptotics approximate the case where people are substantially more likely to match with people who are "close" to them than people who are far away: the ratio $\mathbb{P}[W_{ij} = 1 | \|X_i - X_j\| < C] / \mathbb{P}[W_{ij} = 1 | \|X_i - X_j\| > C]$ diverges as n gets large.

2.3.2 Counterfactual selection

Under a strongly homophilic network formation process, it is possible to derive estimators whose bias asymptotically disappears, even if some confounders are unobserved. Given a group of counterfactuals $\mathcal{C}_i$ for individual i, I define a CATE estimator as

$$\widehat{\text{CATE}}(x_i; \mathcal{C}_i) \stackrel{\text{def}}{=} \frac{1}{|\mathcal{C}_{i1}|} \sum_{j \in \mathcal{C}_{i1}} Y_j(T_j) - \frac{1}{|\mathcal{C}_{i0}|} \sum_{j \in \mathcal{C}_{i0}} Y_j(T_j) \tag{2}$$

where $\mathcal{C}_{it} \stackrel{\text{def}}{=} \mathcal{C}_i \cap \{j | T_j = t\}$. Note that the group of counterfactuals will typically depend on sample size, although the dependence is left implicit in the notation.

A first idea is to rely on friends to construct groups of counterfactuals, *i.e.*, setting $\mathcal{C}_i = \mathcal{N}(i)$. This creates a CATE estimator based on the difference between treated friends and non-treated friends. Thanks to triangular inequality relationships and the nature of homophily, however, one can use further information from the friendship network.

For instance, one can consider friends of friends or higher-order friendships:

$$\begin{aligned} \widehat{\text{CATE}}(x_i; \cup_{m=1}^{M} \mathcal{N}^m(i)) = & \frac{1}{|\cup_{m=1}^{M} \mathcal{N}_1^m(i)|} \sum_{j \in \cup_{m=1}^{M} \mathcal{N}_1^m(i)} Y_j(T_j) \\ & - \frac{1}{|\cup_{m=1}^{M} \mathcal{N}_0^m(i)|} \sum_{j \in \cup_{m=1}^{M} \mathcal{N}_0^m(i)} Y_j(T_j) \end{aligned} \tag{3}$$

for some highest order of friendship M. For $M = 1$, this is the simple estimator that compares treated friends and non-treated friends.

Although this estimator helps reduce the variance by averaging over more observations, it may be more biased and less appealing in larger samples. An alternative estimator relies on friends in common to improve counterfactual quality:

$$\begin{aligned}\widehat{\text{CATE}}(x_i; \{j|\ |\mathcal{N}(i;j)|> \tau\}) = &\frac{1}{|\{j|\ |\mathcal{N}_1(i;j)|\geq \tau\}|} \sum_{j\in\{j|\ |\mathcal{N}_1(i;j)|\geq\tau\}} Y_j(T_j) \\ &- \frac{1}{|\{j|\ |\mathcal{N}_0(i;j)|\geq \tau\}|} \sum_{j\in\{j|\ |\mathcal{N}_0(i;j)|\geq\tau\}} Y_j(T_j)\end{aligned} \tag{4}$$

for some minimum number of common connections τ.

As an illustration, consider Figure 1 where the group of counterfactuals for individual i with unobserved characteristics $x_i \in \mathbb{R}^2$ is depicted in red and the remaining observations in black. The leftmost picture is the infeasible[4] estimator where the researcher is to select all observations below a certain distance. Next on the right, friends are used as the group of counterfactuals, providing a noisy version of (i): selected observations tend to fall close to x_i, but some close observations are unselected while observations farther away may be selected nevertheless.

In the third picture, one can see the effect of picking friends of friends. This allows us to make use of more observations, which will reduce the variance of the estimator, but there may be a tendency to grab more observations outside the sphere too. Finally, the last picture selects individuals who have at least 2 friends in common with i. This typically reduces the bias compared to the previous estimator but selects fewer observations.

2.3.3 Asymptotics

Under *strong homophily*, standard asymptotics are directly achievable with estimators such as (3) or (4). In what follows, I leave the vector of covariates x undecomposed as manipulations on observables are not needed to secure consistency and asymptotic normality. The main possible benefit of observing some covariate is a bias reduction, which is discussed in the Bias management subsubsection. The other main usage of observables is to use truncation to manage networks that are not strongly homophilic; a detailed discussion is the object of Subsection 2.4.

Strong homophilic behavior will ensure that the bias of the proposed CATE estimators disappears. As for the variance, its collapse hinges on averaging over

[4]Except in the extreme network formation process in which individuals select friends deterministically conditional on covariates: $w(x) = \mathbb{1}_{x>C}$. Then, the first two pictures become identical.

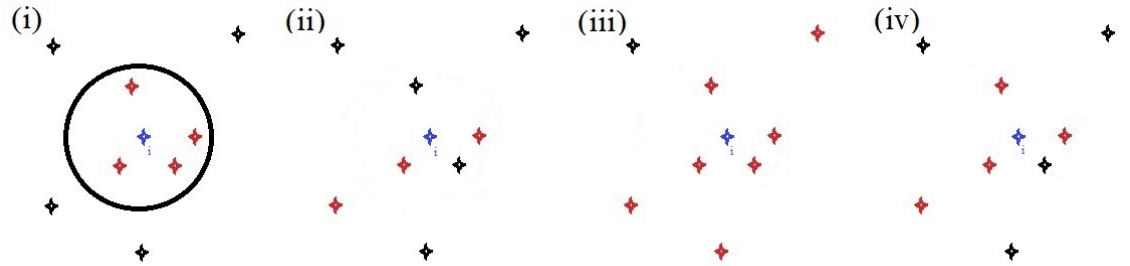

Figure 1: Comparison among possible groups of counterfactuals for individual i. Observations are represented by stars (blue = i; red = selected as counterfactual; black = not selected as counterfactual). From left to right, (i) (Infeasible) individuals within a given distance of x_i, (ii) individuals who are friends with i, (iii) individuals who are friend with i or friend of a friend of i, (iv) individuals who have 2 friends in common with i.

increasingly many observations. The asymptotic behavior of the number of counterfactuals depends on the speed of convergence of w_n, of which the following lemma provides a formal analysis.

Lemma 2.1. *The number of connections for i exceeds any real number with probability approaching one if* $\lim_{n\to\infty} n \int_{\mathcal{X}}(1 - w_n(\|x_j - x_i\|))f_X(x_j)\, dx_j = \infty$.
Under regular strong homophily, this holds, and moreover
a) If $w_n \leq C$ for all sufficiently large n, the probability of forming a connection of order up to M is at least $O\left({s_n}^{-d}\left(n{s_n}^{-d}\right)^{M-1}\right)$. The number of connections for i of order up to M exceeds any real number with probability approaching one.
b) The probability of i and j having at least τ friends in common is at least $O({s_n}^{-d})$. The number of individuals with whom i has at least τ friends in common exceeds any real number with probability approaching one.

The lemma, proven in the appendix, provides conditions under which the number of connections grows to infinity. It also specifies the rates at which the probabilities of forming a connection up to the M-th or having τ friends in common decrease under *strong homophily*, and when all corresponding counts go to infinity. With overlap, the lemma ensures the number of treated and untreated connections both grow to infinity.

The condition in the lemma states that w_n must not increase too fast to ensure that connections are still being formed. Basically, people must not become pickier at too high a rate for their friend count to keep increasing. The formal criterion analyses the integral $\int_{\mathcal{X}}(1 - w_n(\|x\|))f_X(x + x_i)\, dx$, suggesting that link functions

that do not depend on network size or grow uniformly slowly enough such as $w_n(x) = 1 - {s_n}^{-1} g(x)$ induce unbounded friend counts.

Regular strong homophily is not only consistent with a growing count but also further implies a matching quality improvement that is absent when w_n is constant or grows uniformly. Specifically, a sequence such as $w_n(x) = 1 - {s_n}^{-1} g(x)$ would stabilize the 'posterior' distribution $f_{X_j|j\in\mathcal{N}(i)}$; it does not imply that friends are closer on average in larger networks.

As a result, *regular strong homophily* will be key in securing consistency properties. An important part in establishing these is the analysis of the bias

$$\begin{aligned}\mathbb{B}_i \overset{\text{def}}{=} &\mathbb{E}[Y_j(1)|j \in \mathcal{C}_i, T_j = 1]] - \mathbb{E}[Y_j(1)|X_j = x_i] \\ &- (\mathbb{E}[Y_j(0)|j \in \mathcal{C}_i, T_j = 0] - \mathbb{E}[Y_j(0)|X_j = x_i])\end{aligned}$$

which will be shown to disappear under various conditions depending on the group of counterfactuals. Specifically, the following conditions on w_n will ensure that the bias vanishes.

Assumption 2.2 (Hölder continuity of CATE and convergence of link function)**.** a) CATE(x) is Hölder continuous with exponent α on a neighborhood of x_i, *i.e.* for any x, y in the neighborhood $\|\text{CATE}(y) - \text{CATE}(x)\| \leq C\|y - x\|^\alpha$ for some $\alpha > 0$. b) For some $\varepsilon_n \downarrow 0$, either $\mathcal{C}_i = \cup_{m=1}^M \mathcal{N}^m(i)$ and $\sum_{m=1}^M (1 - w_n(\frac{\varepsilon_n}{m}))^m = o({s_n}^{-d})$, or $\mathcal{C}_i = \{j|\ |\mathcal{N}(i;j)| \geq \tau\}$ and $(1 - w_n(\frac{\varepsilon_n}{2}))^\tau = o({s_n}^{-d})$.

Hölder continuity is a standard assumption that imposes a mild degree of smoothness in the CATE. Part b) of the assumption restricts the way w_n converges to 1; it requires a sufficiently fast convergence away from the origin.

The second part of the next theorem obtains a clean rate of $O({s_n}^{-2})$, valid even for the simple first-order friend group of counterfactuals. The basic requirement smoothness of the underlying functions of x, in particular existence of derivatives. Specifically,

Assumption 2.3 (Existence of Derivatives)**.** The conditional expectation $\mathbb{E}[Y_i(t)|X_i = x]$, the propensity score $\mathbb{P}[T_i = 1|X_i = x]$, and the density $f_X(x)$ are differentiable.

Now, the consistency theorem reads

Theorem 2.1 (Consistency)**.** *a) Suppose* $\lim_{n\to\infty} n \int_{\mathcal{X}} (1 - w_n(\|x\|)) f_X(x + x_i)\, dx = \infty$, *and* $\mathbb{E}[Y_j(t)^2] < \infty$ *for* $t = 0, 1$.
Then, $\widehat{\text{CATE}}(x_i; \mathcal{C}_i)$ *is consistent for* CATE(x_i) *under Assumption 1. Moreover, the bias satisfies* $\mathbb{B}_i = O(\varepsilon_n^\alpha + {s_n}^d R)$ *with* $R = \sum_{m=1}^M (1 - w_n(\frac{\varepsilon_n}{m}))^m$ *and* $R = (1 - w_n(\varepsilon_n/2))^\tau$, *respectively, for* $\varepsilon_n \downarrow 0$ *as in Assumption 1.*
b) If network formation is regularly strongly homophilic, $\mathcal{X} = \Omega$, *and Existence of derivatives hold, then the estimator that simply uses friends, i.e.* $\widehat{\text{CATE}}(x_i; \mathcal{N}(i))$, *is consistent with bias* $\mathbb{B}_i = O({s_n}^{-2})$.

The theorem is proven in the appendix. The main difficulty in part a) is to derive an expression for the bias that can subsequently be bounded via homophilic assumptions and triangular inequalities. In the second part, the existence of derivatives allows the use of Taylor expansions so that the derivation shares similarities with nonparametric kernel analysis, though the noisy matching through w_n renders the problem non-standard.

An intuition for the nature and order of the bias can be gathered from the (infeasible) kernel estimator averaging observations within a shrinking sphere centered at x_i. The present estimators constitute noisy versions of the kernel estimator with ${s_n}^{-1}$ playing a role analogous to that of a bandwidth; their biases vanish at a similar rate - possibly slightly more slowly depending on the counterfactual selection strategy and the underlying smoothness.

To sum up, consistency is secured provided homophilic behavior becomes relatively more pronounced and w_n does not grow too fast with sample size. If w_n does not increase to one (so that there is still an asymptotic bias due to different covariates) or does so too quickly (so that the friend count stays bounded), the estimator is no longer consistent. These situations may have empirical relevance and are best addressed with the method in Subsection 2.4.

Before that, I finalize the current analysis with an asymptotic normality result: CATE estimators are asymptotically normal at all x_i under the assumptions for consistency. Formally,

Theorem 2.2 (Asymptotic Normality)**.** *Suppose the Consistency assumptions hold. Then, for a bias* $\mathbb{B}_i$ *at location* x_i*,*

$$\sqrt{|\mathcal{C}_i|}(\widehat{\text{CATE}}(x_i;\mathcal{C}_i) - \text{CATE}(x_i) - \mathbb{B}_i) \to^d \mathcal{N}(0;V)$$

where $V = \frac{\mathbb{V}[Y_j(1)|X_j=x_i]}{\mathbb{P}[T_j=1|X_j=x_i]^2} + \frac{\mathbb{V}[Y_j(0)|X_j=x_i]}{\mathbb{P}[T_j=0|X_j=x_i]^2} - 2\frac{\mathbb{C}[Y_j(1);Y_j(0)|X_i=x_i]}{\mathbb{P}[T_j=1|X_j=x_i]\mathbb{P}[T_j=0|X_j=x_i]}$.
Moreover, $\mathbb{B}_i$ *is asymptotically negligible if* $\text{CATE}(x)$ *is Hölder continuous with exponent* α *on a neighborhood of* x_i *and one of the following holds:*
(i) $\mathcal{C}_i = \cup_{m=1}^{M}\mathcal{N}^m(i)$ *with* $\sum_{m=1}^{M}(1 - w_n(\frac{n^{-\gamma}n}{m}))^m = o({\lambda_n}^2\frac{M}{n})$ *for* $\gamma > \frac{1}{2\alpha}$ *or*
(ii) $\mathcal{C}_i = \{j|\ |\mathcal{N}(i;j)| \geq \tau\}$ *and* ${s_n}^d(1 - w_n(n^{-\gamma}))^\tau = o({\lambda_n}^{-1/2})$ *and* $\gamma > \frac{1}{2\alpha}$ *or*
(iii) Network formation is regularly strongly homophilic, Existence of Derivatives hold, $\mathcal{X} = \Omega$*, and* $\frac{\sqrt{\lambda_n}}{{s_n}^2} \to 0$ *for* $\lambda_n \overset{\text{def}}{=} ns_n^{-d}$.

The proof of the theorem is given in the appendix and is mostly a consequence of the consistency proof, with an analysis of the bias convergence rate.

The size of the group of counterfactuals grows at the rate λ_n. If λ_n grows too fast, the estimator is still consistent but asymptotic normality requires handling the bias, which is the object of the next subsection. Once the bias has been accounted for,

the variance can be estimated with the same kind of truncation methods[5], allowing for statistical inference.

Finally, an important consequence of consistent CATE estimation at any x_i in the presence of unobserved confounders is that the average treatment effect can also be estimated under unobservable-robust unconfoundedness.

2.3.4 ATE inference

Given the last two theorems, one obtains an estimator $\widehat{\text{ATE}}$ by averaging over a CATE estimator at all x_i. The resulting ATE estimator is then consistent and asymptotically normal (possibly upon bias management) under the theorems' conditions and regularity conditions.

Specifically, consider $\frac{1}{n}\sum_{i=1}^{n}\widehat{\text{CATE}}(x_i, \mathcal{C}_i)$ and collect the terms involving $Y_i(T_i)$ for each i to obtain $\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_1(T_i)\sum_{j\in\mathcal{C}_i}\frac{1}{|\mathcal{C}_{j1}|} - \mathbb{1}_0(T_i)\sum_{j\in\mathcal{C}_i}\frac{1}{|\mathcal{C}_{j0}|}\right)Y_i(T_i)$[6]. The weights are well-defined as long as $\mathcal{C}_{jt}$ is non-empty, which occurs with probability approaching one. By including j into its group of counterfactuals and increasing the count $|\mathcal{C}_{j(1-T_j)}|$ by one unit, the weights are also well-defined in finite samples and the estimator will have appropriate asymptotic behavior.

Formally, this leads to the estimator

$$\widehat{\text{ATE}} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}(\hat{\omega}_{i1} - \hat{\omega}_{i0})Y_i(T_i) \tag{5}$$

where $\hat{\omega}_{it} \stackrel{\text{def}}{=} \mathbb{1}_t(T_i)\sum_{j\in\mathcal{C}_i}\frac{1}{1+|\mathcal{C}_{jt}|}$. Its asymptotic distribution is described by the next theorem:

Theorem 2.3. *Suppose $\mathbb{E}[Y_j(t)^2] < \infty$, condition (iii) of Theorem 2.2 holds, $\frac{\sqrt{n}}{{s_n}^2} \to 0$, for $t \in \{0,1\}$, and the density's tails are convex (i.e., there exists a compact K such that $f_X(x)$ is convex on K^c). Then,*

$$\sqrt{n}(\widehat{\text{ATE}} - ATE) \to^d N(0; \mathbb{V}[\omega_i Y_i(T_i)])$$

where $\omega_i \stackrel{\text{def}}{=} \omega_{i1} - \omega_{i0}$, $\omega_{it} \stackrel{\text{def}}{=} \frac{\mathbb{1}_t(T_i)}{\mathbb{P}[T_i=t|X_i]}$ for $t \in \{0,1\}$.

The theorem is proven in the appendix and requires a careful decomposition and bounding of the deviations of the $\hat{\omega}$'s from their true value. The addition of one

[5]The covariance, as conditioned on $X_j = x_i$, vanishes under zero correlation in the corresponding 'error terms'. Often, variables affecting both $Y(1)$ and $Y(0)$ induce a positive correlation in errors so that assuming zero covariance leads to conservative inference

[6]This last expression holds as long as $i \in \mathcal{C}_i$ implies $j \in \mathcal{C}_i$, which is true for the groups of counterfactuals previously discussed.

unit in the denominators of $\hat{\omega}_{it} = \mathbb{1}_t(T_i) \sum_{j \in \mathcal{C}_i} \frac{1}{1+|\mathcal{C}_{jt}|}$ proves useful in improving the centering of the terms, thereby lowering the bias.

The convex tail condition is an uncommon but minor regularity condition on the asymptotic behavior of the density. It plays a role in deriving a bound allowing the use of the dominated convergence theorem and can likely be further relaxed.

This theorem allows one to conduct inference about the average treatment effect under (unobservable-robust) unconfoundedness. The variance can be estimated using sample averages and the estimated weights.

2.3.5 Bias management

I briefly outline various strategies to deal with such a fast growth of λ_n that the bias affects the asymptotics.

First, note that the analysis focuses on continuous random variables, which are responsible for asymptotic biases in nonparametric estimation. If some of the variables are discrete, they do not contribute to potential bias issues and dimensionality can be reduced accordingly.

If some variables are observed, it is possible to use more conventional estimators on the observed side and to truncate the sums involved in the construction of the corresponding estimators to account for unobservables. The use of parametric or high-order kernels type of estimators for the observed part allows one to relegate bias issues to the unobserved world. Some of these possibilities are explored in Appendix B.

As a result, the bias mostly creates difficulties when unobservables are high-dimensional. There are two main possibilities to deal with this issue: attempting a (noisy) smoothing along unobserved dimensions and approximating the bias.

The use of connections in common may allow for further smoothing, albeit of a noisy type, by using weighted averages (weight τ_{ij})/$\sum_j \tau_{ij}$) or kernels such as $\frac{\sum_j K((x_j^o - x_i^o)/h; 1/\tau_{ij}) y_j}{\sum_j K((x_j^o - x_i^o)/h; 1/\tau_{ij})}$, where τ_{ij} counts common friends. Nevertheless, the properties of the latter estimator are unknown and simulations indicate that simple weighted averages typically perform better. A more refined way to perform smoothing is available via the estimators developed in the next subsection so that the discussion on unobservable smoothing is deferred to Subsection 2.4.

Alternatively, one can try and obtain an approximation or a bound on the bias to perform inference. A way to capture the magnitude of the bias is to specify (bounds on) partial effects, distributions, and link formation.

For instance, suppose f_X is normal with mean μ and variance $\Sigma = H^{-1}$, while $w_n(\|x\|) = 1 - e^{-s_n \|x\|^2/2}$ (it often makes sense to standardize the vector X for such

network formation models). The numerator of the 'posterior' distribution of X_j given a friendship link with i reads
$(2\pi)^{d/2}|\Sigma|^{-1/2}e^{-1/2(x-\mu)'H(x-\mu)}e^{-1/2s_n\|x-x_i\|^2}$ Re-arranging, we obtain
$(2\pi)^{d/2}|\Sigma|^{-1/2}e^{-1/2(x-\overline{\mu})'\overline{H}(x-\overline{\mu})}e^{-1/2(x_i-\mu)'(H-H\overline{H}^{-1}H)(x_i-\mu)}$ where $\overline{H} \stackrel{\text{def}}{=} H + s_n I$; $\overline{\mu} \stackrel{\text{def}}{=} x_i - \overline{H}^{-1}H(x_i - \mu)$ Therefore, the posterior distribution is normal with mean $\overline{\mu}$ and variance $\overline{H}^{-1}$.

Remark that obtaining a closed-form solution typically hinges on specifying a link formation w that combines well with the underlying density f_X, which is algebraically equivalent to the use of conjugate priors in Bayesian statistics.

Specifying a bound on the CATE then delivers an approximation to the bias, which allows one to perform inference by extending the normal confidence intervals of Theorem 2.2. In practice, it may be of interest to assess how large the Hölder constant, how fat the tails of unobservable distributions, or how large the derivatives of underlying functions would have to be for the confidence interval to include 0. One can also get a sense of the range of possible bias values by varying inputs and then discussing resulting confidence intervals.

2.4 Homophily in unobservables

Now, I consider a link formation model that need not depend on sample size and allow for non-homophilic behavior along observed covariates. Intuitively, this framework is more suitable when it is believed that the strength of homophilic selection is lower than desired, so that one wants to lower the bias of estimators relying on connections as counterfactuals, and when individuals are believed to match non-homophilically on some observed covariates.

Compared to the previous framework, the estimator developed in this subsection is intuitively more useful when bias is the main concern. For instance, if the number of peers is relatively high, there are covariates known to affect network links but not the outcome equation, or there are more concerns about nonlinearities in treatment effects than about noise.

Now, links are formed as $W_{ij} = \mathbb{1}_{w(h(X_i^o;X_j^o)+\|X_i^u - X_j^u\|)\leq \eta_{ij}}$. The function w does not explicitly depend on n anymore, though some dependence could be accounted for. The function h need not be homophilic nor separable in the observed X^o but is assumed known. Most results would apply with minor modifications if a lower bound with the relevant properties can be obtained. The dimension of $\mathcal{X}^u$ is denoted by d_u.

If the observables are relevant to the outcome equation, they can be controlled for with standard methods. For instance, one might restrict the sample to the j's whose distance $\|X_j^o - x_i^o\|$ is low enough or use bias-reduction techniques such as

kernel smoothing. To get the main point, I will abstract from observables in the outcome equation and illustrate how to control for the unobserved components. One can consider the outcomes Y in what follows as a pre-controlled version, for instance with $Y = \mathbb{1}_{\|X^o - x_i^o\| < \varepsilon}\tilde{Y}$ or $Y = K((X^o - x_i^o)/b)/(\sum_j K((X_j^o - x_i^o)/b))\tilde{Y}$ where $\tilde{Y}$ is from the original sample. In the former case of simple truncation, it is straightforward to adapt the results.

I consider $\frac{1}{|\mathcal{C}_{i1}(\kappa)|}\sum_{j\in\mathcal{C}_{i1}(\kappa)} Y_j(T_j) - \frac{1}{|\mathcal{C}_{i0}(\kappa)|}\sum_{j\in\mathcal{C}_{i0}(\kappa)} Y_j(T_j)$, where the group of counterfactuals now depends on a truncation parameter κ. It will play a key role by placing a lower bound on h, inducing a closer distribution of unobservables among friends.

The main idea is that if there is no observed rationale for two people being friends, the reason for their friendship likely lies in the unobserved world. Then in the present model[7], people that are friends despite a high value of h are less likely to differ strongly on unobservables. To see this, consider the case where people reject friendship with anyone whose quality of match is too poor (*i.e.*, $w(x) = 1$ for x large enough). Then, two friends with an h close to the boundary must have close unobservables since a high discrepancy in unobservables would have brought $h + \|X_i^u - X_j^u\|$ above the threshold.

Specifically, I consider $\mathcal{C}_i(\kappa) \stackrel{\text{def}}{=} \{j \in \mathcal{N}(i) | h(x_i^o, x_j^o) > \kappa\}$. The estimator then truncates the sums to select individuals whose observed characteristics make them unlikely to be friends. κ is viewed as a sequence converging to ∞ at a rate to be determined.

Using this strategy, a counterpart to Theorem 2.2 can be derived with κ-truncation replacing *strong homophily*.

Theorem 2.4. *Suppose $CATE(x)$ is Hölder continuous with exponent α on a neighborhood of x_i, $1 - w$ has bounded support, and there exists a sequence b_n such that κ-truncation satisfies $(1 - w(\kappa + b_n))b_n^{d_u+1} = O({s_n}^{-d})$ and a sequence $\varepsilon_n \downarrow 0$ satisfying $\kappa + \varepsilon_n > \sup\{supp\{1\text{-}w\}\}$ eventually and $\sqrt{\lambda_n}{\varepsilon_n}^{\alpha}$.*
Then, the estimator $\widehat{\text{CATE}}(x_i; \mathcal{C}_i(\kappa))$ satisfies

$$\sqrt{|\mathcal{C}_i|}(\widehat{\text{CATE}}(x_i; \mathcal{C}_i(\kappa)) - \text{CATE}(x_i)) \to^d \mathcal{N}(0; V)$$

The condition $(1 - w(\kappa + b_n))b_n^{d_u+1} = O({s_n}^{-d})$ restricts the speed at which κ can increase so that the number of observations used in estimating the CATE keeps growing. The first two terms pertain to the behavior of the w function; the second pertains to the space in which unobservables live.

[7] One could imagine variations on the network formation process, where this intuition would be rationalized slightly differently, leading to another estimator. One instance is briefly discussed in appendix B.

The term $1-w(\kappa+b_n)$ comes from the increasing cost of truncating as potential connections are accepted at decreasing rates. In the presence of a discontinuity at the end of the support, *i.e.*, $w(x) = a + (1-a)\mathbb{1}_{x>D}$ for $a \in [0;1[$ and $D \in \mathbb{R}^+$, this term disappears. The second term, ${b_n}^{d_u+1}$, is the result of forcing unobservables in a b_n-ball using values of h lying between κ and $\kappa + b_n$.

A natural estimator of the end of support, if unknown, is the highest value of h among all i, j satisfying $W_{ij} = 1$. One can extend the theorem to unbounded support of $1-w$, provided $1-w$ decreases sufficiently quickly (in practice, $1-w = e^{-x^2}$ provides a sufficiently fast decay). In this case, a term $\rho_h(\kappa + b_n/2)$, where ρ_h is (a lower bound on) the tail decay of h conditional on $X_j^u \in B_r(x_i^u)$ for some r, arises. The reason is the tail decay of the density while one seeks increasingly larger values of h.

The conditions on ε_n ensure that the bias disappears sufficiently fast to allow standard inference. A more primitive statement is $(1-w(\kappa+\varepsilon_n)) = o({s_n}^{-d})$, which mirrors conditions such as those in Assumption 1, part b), but one can focus on $\kappa + \varepsilon_n$ eventually crossing the end of the support of $1-w$ if it is finite.

Again, one can consider variations on the above scheme, for instance by including friends of friends or truncating based on the number of friends in common.

A possible concern about the previous estimator is that h may feature homophilic behavior as well, in which case there is a conflict between requiring closeness in observables (to improve counterfactual quality) and difference in observables (to induce closeness in unobservables) at once.

The issue can be addressed in various ways. Often, there are variables excluded from the outcome equations that induce the necessary variation in h. A common example is physical distance, which typically affects the likelihood of friendship but not the outcome.

Alternatively, it is possible to work with 'triangles' of friends to control the overall distance (in terms of both observables and unobservables) directly. Compared to the previous κ-truncation, this strategy allows selecting observations whose values of h are low.

Adapt the group of counterfactuals $\mathcal{C}_i(\kappa)$ to be defined as
$\{j \in \mathcal{N}(i) | \exists k \in \mathcal{N}(i) \cap \mathcal{N}(j), h_{ik} > \kappa, h_{jk} > \kappa\}$. Members of this group are close in observables by selection and more likely to be close in unobservable because they have a common friend with i with relatively high observed differences. Formally,

Theorem 2.5. *The estimator* $\widehat{\mathrm{CATE}}(x_i; \mathcal{C}_i(\kappa))$ *satisfies*

$$\sqrt{|\mathcal{C}_i|}(\widehat{\mathrm{CATE}}(x_i; \mathcal{C}_i(\kappa)) - \mathrm{CATE}(x_i)) \to^d \mathcal{N}(0; V)$$

if $CATE(x)$ is Hölder continuous with exponent α on a neighborhood of x_i, h is homophilic, and the sequence κ satisfies $((1-w(\kappa+b_n))b_n^{d_u+1}\rho_h(\kappa+b_n/2))^2 = O(\frac{\lambda_n}{n^2})$ and $(1-w(\kappa+\varepsilon))(1-w(\kappa)) = o(\frac{\lambda_n}{n^2})$ where ρ_h is (a lower bound on) the tail decay of h conditional on $X_j^u \in B_r(x_i^u)$ for some r.

Note that relying on triangles of friends also provides further opportunities in terms of bias management, in case λ_n grows too fast for the bias to remain asymptotically negligible. Indeed, it is possible to smooth along both observed and unobserved dimensions by using the (noisy) measure of distance in the covariates suggested by triangular inequalities.

3 Simulations

I assess the performance of the estimators through simulations. I consider various outcome equations and network formation processes and measure the resulting root mean square error (RMSE). For comparison, the RMSE of standard estimators – using available observed variables – is also provided. Specifically, ordinary least squares and kernel regression on observables will be used as benchmarks.

Variables are generated as follows: a random vector V, whose components are one student with five degrees of freedom and two normally distributed variables (all normalized at unit variance), is used to construct

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 0.7 & 0.3 & 0 \\ -0.1 & 1 & 0.4 \\ 0 & -0.6 & 0.7 \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix}$$

so that there is a non-negligible correlation structure among the components of X. In particular $\text{corr}(X_1, X_2) = 0.28$, $\text{corr}(X_1, X_3) = -0.26$, and $\text{corr}(X_2, X_3) = -0.32$. The first two variables are observed, but the last one is not, *i.e.*, $X^o = (X_1, X_2)$ and $X^u = X_3$.

The propensity score follows a logistic distribution with argument $X\beta$, where $\beta = \begin{pmatrix} 0.5 & 0.5 & \beta_3 \end{pmatrix}'$, and treatment status is then drawn conditional on X. This ensures the propensity score varies widely, often ranging between 0.05 and 0.95. The parameter β_3 controls the selection on unobservables. Three values will be considered: $0, 0.5, 1$. In the first case, the probability of being treated does not change with X_3; the second case puts the unobserved variable on a similar footing as the observed ones while the last value reflects a setting in which the most relevant covariate was unavailable. Thus, the performance of traditional methods that cannot account for the unobserved component is expected to deteriorate as β_3 increases.

The outcome equation is determined by

$$Y_{Ai} = 1 + (e^{X_{1i}/5} + e^{X_{2i}/5})\, T_i + X_{1i} + X_{2i} + X_{3i} + X_{1i}X_{2i} + \varepsilon_i$$

or

$$Y_{Bi} = 1 + 4\Phi((-X_{i1} + X_{i2} + X_{i3})/2)\ T_i + X_{1i} + X_{2i} + X_{3i} \\ + X_{1i}X_{2i} + (X_{1i} + X_{2i})X_{3i} + \varepsilon_i$$

where ε_i is normally distributed. The first equation is a linear model in the unobservables with a treatment effect heterogeneity only due to observables; the second version has a stronger treatment effect heterogeneity and nonlinear interactions, both of which are affected by the unobserved variable. The first specification should be favorable to standard methods, especially for low values of β_3, while the second specification is expected to create significant bias.

The network formation processes are $w_a(x) = \mathbb{1}_{x<D}$, $w_b(x) = \Phi(x - D)$, $w_c(x) = \min\{x/D, 1\}$, and $w_d(x) = 1 - e^{-\frac{1}{2D}x^2}$. The sample size is $N = 500$ and D is calibrated so that the average number of friends is roughly four as in the application (this corresponds to $D = 14, 15, 25$, and 1.17, respectively). I consider the *strong homophily* case and scale the argument of the network function by s_n: $w_n(x) = w(s_n x)$ with $s_n = N/\ln(N)$).

The first function is simple and unlikely to represent an exact network formation process. It is however a useful benchmark: it constitutes an 'ideal' (noiseless) kernel smoothing scenario in which the network provides a deterministic truncation conditional on covariates. Furthermore, it can also be viewed as an approximation to a smooth function with a steep increase around a critical threshold. This could occur for instance when people are not allowed to, have limited possibility to, or do not wish to form ties outside of their community or some defining characteristics, or when the costs of forming friendships start rising sharply outside some comfort zone.

The second function is smooth but rises sharply around point D. The third function is a continuous function with bounded support for $1 - w$, while the fourth is a smooth function that reaches 1 only asymptotically.

I first discuss results pertaining to the estimation of the ATE, which is often the parameter of interest in empirical studies, then provide results for CATE estimation. The full results for the ATE are displayed in appendix C, detailing the performance of the estimators as M or τ vary. For easier review, I report the performance of the most relevant specifications $M = 3$, $\tau = 1$, and OLS in Table 1 below.

OLS generally performs well when there is no selection on unobservables but its performance quickly deteriorates as β_3 increases. By contrast, the estimators developed in the paper do not suffer so much and tend to dominate OLS across specifications in terms of ATE estimation. In the presence of unobserved confounding, the properties of OLS become quickly unappealing, especially when the unobserved variable has additional non-linear impacts on the outcome equation as in scenario B.

Table 1: Performance (in terms of RMSE for ATE) of estimators

β_3	0								0.5			
y	A				B				A			
w	a	b	c	d	a	b	c	d	a	b	c	d
$M=3$	0.13	0.12	0.13	0.15	0.14	0.12	0.13	0.14	0.12	0.15	0.18	0.20
$\tau=1$	0.15	0.14	0.13	0.12	0.17	0.15	0.14	0.14	0.13	0.12	0.12	0.12
OLS	0.15				0.18				0.37			
β_3	0.5				1							
y	B				A				B			
w	a	b	c	d	a	b	c	d	a	b	c	d
$M=3$	0.12	0.15	0.20	0.22	0.13	0.17	0.21	0.24	0.18	0.21	0.28	0.30
$\tau=1$	0.12	0.12	0.13	0.14	0.12	0.13	0.13	0.14	0.13	0.14	0.17	0.18
OLS	0.48				0.63				0.86			

RMSE of the estimator using friends up to the third order, the estimator using individuals with at least one friend in common, and OLS.

In the CATE case, one has to select an individual whose treatment effect is of interest. I focus on an individual with average observed characteristics[8] ($X^o = 0$ and X^u drawn conditioning on $X^o = 0$) as to target $\mathbb{E}[Y_i(1) - Y_i(0)|X_i^o = 0]$.

The results are displayed in Appendix C. Standard methods perform particularly well with outcome equation A and $\beta_3 = 0$. In this case, there is no selection on unobservables and the treatment effect varies only according to observed characteristics. If one of these elements is modified, however, the performance deteriorates. Moreover, the kernel estimator generally struggles because of the unobserved variable – even when it affects only the outcome and not selection into treatment.

By contrast, the estimators that leverage network information are able to perform similarly across specifications, regardless of the strength of selection in unobservables. Here, the ability to average over a larger number of observations is particularly valuable because the number of friends is small and one targets the CATE. This explains why larger values of M and lower values of τ perform better: when the number of peers tends to be small, the estimator benefits from lowering the variance. This also explains the better performance of the higher-order friend

[8] Alternative target individuals could be (i) an individual with average characteristics, including unobservables, or (ii) randomly chosen individuals. Although they lead to similar results, the latter is less meaningful here since the friend count drops frequently to 0 for individuals farther from the norm. I provide the results from randomly sampling the individual of reference for a larger dataset of 1,000 in the appendix.

version of the estimator. Overall, using friends of order up to 3 tends to deliver the best results and to dominate standard methods that use only observables except in the first scenario.

The type of network link function (a, b, c, d) appears to have little impact on the performance. To the extent that the estimators can be interpreted as (noisy) kernel estimators, this mirrors the result about the (lack of) influence of the choice of kernel.

4 Application

I apply the method to estimate the effect of parental involvement on students' test scores. I use the dataset from the project "Attitudes and Relationships among Primary and High School Students", see Portella and Kirschbaum (2022). The dataset contains information about 4409 Brazilian students, their beliefs, and friendship ties among them.

The outcome of interest is the average grade, ranging from 0 to 10, and the treatment is the level of parent support. The average is taken over math, Portuguese, English, history, geography, and art grades[9]. For comparability across school years, I normalize the grade by subtracting the mean over students in a given year. The dataset contains a (self-reported) score for parent support which ranges from around -3 to 1; the treatment is dichotomized by truncating around the mean of 0 and the goal is to estimate the ATE.

Although some covariates are available – age, gender, race, religion, whether parents are employed, poverty, and importance of study for the child, they are likely at best proxies for the underlying causes. Thus, omitted variable bias remains a concern, especially due to the usual unobserved ability.

The sign of the omitted variable bias is far from straightforward in this instance, even assuming that the importance of study score controls adequately for effort. Intuitively, children with lower ability may require more attention but are also more likely through heredity to have less able parents, who may be less inclined or able to help.

Formally, a possible model for grade, y_i, would be $y_i = F(a_i; e_i; s_i)$ where F is an unknown function, a_i is ability, e_i is the level of effort or work, and s_i is parent support. Denoting parents' ability by A_i, one could model $a_i = A_i + \text{noise}_i$ and $s_i = \delta_1 a_i + \delta_2 A_i + \text{noise}_i$ with independent noises. This would imply $s_i = (\delta_1 + \delta_2)a_i + \text{noise}_i$, where the signs of the deltas are likely to be negative and positive, respectively, leading to an indeterminate sign for the correlation. Hence, it

[9] Although additional subjects are available, including them would require dropping a substantial number of observations because of missing data.

is not obvious whether parents pay more attention to children with higher or lower ability. Moreover, the function F is unlikely to be linear since, *e.g.*, ability may increase the return to effort and parent support and there may be nonlinear returns to efforts and/or ability.

As a result, not only is the OLS estimate likely to be unreliable, but it is also hard to figure out the direction of the bias. This motivates the use of alternative methods that can account for the presence of unobservables.

There is evidence that people, teenagers in particular, sort on intelligence. Some studies (Clark and Ayers, 1992; Burgess et al., 2011; Boutwell, Meldrum and Petkovsek, 2017) have documented homophilic matching on various measures of intelligence among teenagers. According to Boutwell, Meldrum and Petkovsek (2017), "preadolescent friendship dyads are robustly correlated on measures of general intelligence".

It is thus plausible that the high school students in the sample have accounted for ability when forming friendship ties, suggesting an avenue for correcting the ability bias through the estimator developed in this paper.

There is some evidence for homophilic matching on some of the observed covariates: age, gender, the importance of study, and poverty. This can be seen in the average distance in the covariate, which is substantially reduced when the average is computed within the peer group[10].

Table 2: Average distance in observables

	Age	Gender	Study score	Poverty score
Average distance (overall)	2.92	0.50	0.96	0.82
Average distance (friends)	0.64	0.26	0.84	0.76

Combined with the small number of peers, which averages four, and the goal of estimating the ATE, this suggests that the *strong homophily* asymptotics developed in section 2.3 provides the relevant framework. With a limited friend count[11], using the higher-order friendship version of the estimator makes intuitive sense and is supported by simulation results.

For each individual, I use the friends up to the third order as the group of counterfactuals, whose average size is fifteen, and compute $\widehat{\text{ATE}}$.

[10]By contrast, the average distance decreases by less than 2% for other variables such as race or religious affiliation, suggesting these variables were not relevant for network formation in this sample.

[11]The number of people with no reported friend is relatively high, but most of these students are also missing covariates. This suggests zero friend counts are more indicative of a missing data problem than of a general asocial behavior. Consequently, I restrict the sample to the sample used for OLS, which also facilitates comparability. The effective sample size is then 2777.

According to OLS, the ATE is 0.06 with no statistical significance at conventional levels (se 0.05). The estimator developed in the paper, however, suggests a higher effect of 0.21 (se 0.05). Because of the ability to control for unobserved ability – including potential nonlinear interactions, the latter estimate may be more reasonable. Going from a low to a high level of parent support increases the mean grade by almost a quarter of a point on average.

Table 3: Estimates and standard errors

	Estimated ATE	standard error
$\widehat{\text{ATE}}$	0.21	0.05
OLS	0.06	0.05

Remark 1: running a regression without the parent's employment status and the poverty score increases the treatment effect estimate from OLS to 0.13. Going back to the model $s_i = \delta_1 a_i + \delta_2 \text{parents' ability}_i + \text{noise}_i$, this may indicate that OLS may be *more* biased upon controlling for parent's employment and poverty because these variables may act as proxies for parent ability and thus increase the conditional correlation between ability and parent support.

Remark 2: if one believes that the importance of study score accurately reflects effort, results in average distance from Table 2 may suggest that the estimate would benefit from a stricter control on this variable. A simple way[12] to address this concern is to estimate the effect on high and low score (dichotomizing at the mean, which is 0) students and then aggregate. This results in a slightly lower average treatment effect estimate of $0.2175\ 1453/2786 + 0.0740\ 1333/2786 = 0.15$.

Remark 3: Switching to $M = 4$ or considering people with at least a friend in common as the group of counterfactuals delivers similar results: the effect is estimated to be 0.21 or 0.17.

References

Aronow, Peter M, and Cyrus Samii. 2017. "Estimating average causal effects under general interference, with application to a social network experiment." *The Annals of Applied Statistics*, 11(4): 1912–1947.

Arpino, Bruno, Luca De Benedictis, and Alessandra Mattei. 2017. "Implementing propensity score matching with network data: the effect of the General

[12] Alternatively, the regression framework developed in appendix B (or even kernel-based methods) could be used. This gives similar results.

Agreement on Tariffs and Trade on bilateral trade." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3): 537–554.

Auerbach, Eric. 2019. "Identification and estimation of a partially linear regression model using network data." *arXiv preprint arXiv:1903.09679*.

Auerbach, Eric. 2022. "Identification and estimation of a partially linear regression model using network data." *Econometrica*, 90(1): 347–365.

Boucher, Vincent. 2015. "Structural homophily." *International Economic Review*, 56(1): 235–264.

Boucher, Vincent, and Ismael Mourifié. 2017. "My friend far, far away: a random field approach to exponential random graph models." *The econometrics journal*, 20(3): S14–S46.

Boutwell, Brian B, Ryan C Meldrum, and Melissa A Petkovsek. 2017. "General intelligence in friendship selection: A study of preadolescent best friend dyads." *Intelligence*, 64: 30–35.

Burgess, Simon, Eleanor Sanderson, Marcela Umaña-Aponte, et al. 2011. *School ties: An analysis of homophily in an adolescent friendship network.* Centre for Market and Public Organisation.

Chao, Min-Te, and WE Strawderman. 1972. "Negative moments of positive random variables." *Journal of the American Statistical Association*, 67(338): 429–431.

Clark, ML, and Marla Ayers. 1992. "Friendship similarity during early adolescence: Gender and racial patterns." *The journal of psychology*, 126(4): 393–405.

Cribari-Neto, Francisco, Nancy Lopes Garcia, and Klaus LP Vasconcellos. 2000. "A note on inverse moments of binomial variates." *Brazilian Review of Econometrics*, 20(2): 269–277.

Currarini, Sergio, Matthew O Jackson, and Paolo Pin. 2009. "An economic model of friendship: Homophily, minorities, and segregation." *Econometrica*, 77(4): 1003–1045.

Demirer, Mert. 2019. "Partial Identification of Linear Models Using Homophily in Network Data." *Working paper*.

De Paula, Aureo. 2017. "Econometrics of network models." Vol. 1, 268–323, Cambridge University Press Cambridge.

Dzemski, Andreas. 2019. "An empirical model of dyadic link formation in a network with unobserved heterogeneity." *Review of Economics and Statistics*, 101(5): 763–776.

Forastiere, Laura, Edoardo M Airoldi, and Fabrizia Mealli. 2020. "Identification and estimation of treatment and interference effects in observational studies on networks." *Journal of the American Statistical Association*, 1–18.

Goldsmith-Pinkham, Paul, and Guido W Imbens. 2013. "Social networks and the identification of peer effects." *Journal of Business & Economic Statistics*, 31(3): 253–264.

Graham, Bryan S. 2014. "An empirical model of network formation: detecting homophily when agents are heterogeneous." *Working paper*.

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

Graham, Bryan S. 2016. "Homophily and transitivity in dynamic network formation." National Bureau of Economic Research.

Imbens, Guido W. 2004. "Nonparametric estimation of average treatment effects under exogeneity: A review." *Review of Economics and statistics*, 86(1): 4–29.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

Imbens, Guido W, and Jeffrey M Wooldridge. 2009. "Recent developments in the econometrics of program evaluation." *Journal of economic literature*, 47(1): 5–86.

Jackson, Matthew O. 2010. *Social and economic networks.* Princeton university press.

Jackson, Matthew O, Zhongjian Lin, and Ning Neil Yu. 2020. "Adjusting for peer-influence in propensity scoring when estimating treatment effects." *Available at SSRN 3522256*.

Johnsson, Ida, and Hyungsik Roger Moon. 2021. "Estimation of peer effects in endogenous social networks: Control function approach." *Review of Economics and Statistics*, 103(2): 328–345.

Liu, Lan, Michael G Hudgens, Bradley Saul, John D Clemens, Mohammad Ali, and Michael E Emch. 2019. "Doubly robust estimation in observational studies with partial interference." *Stat*, 8(1): e214.

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.

Moody, James. 2001. "Race, school integration, and friendship segregation in America." *American journal of Sociology*, 107(3): 679–716.

Newey, KW, and D McFadden. 1994. "Large sample estimation and hypothesis." *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, 2112–2245.

Portella, Alysson, and Charles Kirschbaum. 2022. "Replication Data for: Racial Social Norms among Brazilian Students: Academic Performance, Social Status and Racial Identification."

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, 66(5): 688.

Sánchez-Becerra, Alejandro. 2021. "Spillovers, Homophily, and Selection into Treatment: The Network Propensity Score." *Job market paper*.

Sofrygin, Oleg, and Mark J van der Laan. 2016. "Semi-parametric estimation and inference for the mean outcome of the single time-point intervention in a causally connected population." *Journal of causal inference*, 5(1).

Appendix A: Proofs

4.1 Lemma 3.1

Proof. Consider a sample of $(n+1)$ observations, including i. By independence, $\mathcal{N}(i) = \sum_{j\neq i, j=1}^{n+1} \mathbb{1}_{w_n(\|x_i - X_j\|)\leq \eta_{ij}} \sim \mathcal{B}(n, p_n)$ with
$p_n \overset{\text{def}}{=} 1 - \mathbb{E}[w_n(\|x_i - X_j\|)] = \int_{\mathcal{X}} (1 - w_n(\|x_j - x_i\|)) f_X(x_j)\, dx_j$ by the law of iterated expectation and distributional properties of η. If $np_n \to C > 0$, the friend count asymptotically follows a Poisson distribution with parameter C, while slower sequences p_n induce an unbounded friend count: by Chebyshev inequality we have for any N and large sufficiently large n

$$\begin{aligned}
\mathbb{P}[\mathcal{N}(i) \leq N] &\leq \mathbb{P}[|\mathcal{N}(i) - np_n| \geq np_n - N] \\
&\leq \frac{np_n(1-p_n)}{(np_n - N)^2} \\
&= \frac{1-p_n}{np_n(1-\frac{N}{np_n})^2}
\end{aligned}$$

and thus $\mathbb{P}[\mathcal{N}(i) \leq N] \to 0$ as long as $np_n = n\int_{\mathcal{X}}(1-w_n(\|x_j-x_i\|))f_X(x_j)\,dx_j \to \infty$.

This is the case when the network formation is regularly strongly homophilic. Indeed, consider

$$\begin{aligned}\int_{\mathcal{X}} w_n(\|x-x_i\|)f_X(x)\,dx &= 1-\int_{\mathcal{X}} g(s_n\|x-x_i\|)f_X(x)\,dx\\ &\leq 1-\underline{f}\int_{B_{\underline{r}}(x_i)} g(s_n\|x-x_i\|)\,dx\\ &\leq 1-\frac{\underline{f}}{{s_n}^d}\int_{B_{s_n\underline{r}}(0)} g(\|y\|)\,dy\\ &\leq 1-\frac{\underline{f}}{{s_n}^d}\int_{B_{\underline{r}}(0)} g(\|y\|)\,dy\end{aligned}$$

Since $\int_\Omega g(\|y\|)\,dy > 0$ and g is decreasing, $\int_{B_{\underline{r}}(0)} g(\|y\|)\,dy > 0$ (otherwise, $\int_{B_{\underline{r}s_n}(0)} g(\|y\|)\,dy = 0$ implies $g \equiv 0$, contradicting the former assertion). Thus, using $n{s_n}^{-d} \to 0$, *regular strong homophily* suffices to establish $\lim_{n\to\infty} n\int_{\mathcal{X}}(1-w_n(\|x_j-x_i\|))f_X(x_j)\,dx_j = \infty$.

For part a), write $w_n(\|x_i-x_j\|)$ as w_{ij} and let $\varepsilon > 0$ be small enough for $g(\varepsilon) > 0$ (such an ε exists: otherwise $g = 0$ except possibly at the origin, contradicting $\int_\Omega g(\|y\|)\,dy > 0$).

Then, for n large, j is a friend of m-th order with probability

$$\begin{aligned}&\sum_{j_0,\cdots,j_m}\mathbb{P}[W_{j_kj_{k+1}}=1, W_{j_kj_l}=0\ l\neq k\pm 1]\\ &= (n-1)\cdots(n-m+1)\mathbb{E}[\mathbb{E}[\mathbb{1}\{w_{j_kj_{k+1}}\leq\eta_{j_kj_{k+1}}, w_{j_kj_l}\geq\eta_{j_kj_l},\forall k,l>k+1|\{X_{j_k}\}]]\\ &= \frac{(n-1)!}{(n-m)!}\mathbb{E}\left[\prod_{k=0}^{m-1}(1-w_{j_kj_{k+1}})\prod_{l\neq k\pm1} w_{j_kj_l}\right]\\ &\geq C\frac{(n-1)!}{(n-m)!}\int_{(\mathcal{X})^m}\prod_{k=0}^{m-1} g(s_n\|x_{j_k}-x_{j_{k+1}}\|)\prod_{k=1}^{m} f(x_{j_k})\prod_{l>k+1} w_{j_kj_l}\,d\left(\prod_{k=1}^{m}x_{j_k}\right)\\ &\geq C\frac{(n-1)!}{(n-m)!}\int_{\left(B_{\frac{\min\{\varepsilon,\underline{r}\}}{2s_n}}(x_i)\right)^m} g(\varepsilon)^m\prod_{k=1}^{m} f(x_{j_k})\,d\left(\prod_{k=1}^{m}x_{j_k}\right)\\ &\geq C\frac{(n-1)!}{(n-m)!}\int_{\left(B_{\frac{\min\{\varepsilon,\underline{r}\}}{2s_n}}(x_i)\right)^m} 1\,d\left(\prod_{k=1}^{m}x_{j_k}\right)\\ &= C\frac{(n-1)!}{(n-m)!}s_n^{-md}\end{aligned}$$

Noting that C depends on m only through a factor c^m, $c \stackrel{\text{def}}{=} \underline{f}\underline{w}\frac{\pi^{d/2}}{\Gamma(d/2+1)}g(\varepsilon)$, a lower bound for connections up to order M is obtained from

$$\begin{aligned}\sum_{m=1}^{M} c^m {s_n}^{-dm}\frac{(n-1)!}{(n-m)!} &\geq {s_n}^{-d}\sum_{m=1}^{M}\left(c{s_n}^{-d}(n-M+1)\right)^{m-1}\\ &= {s_n}^{-d}\frac{\left({s_n}^{-d}c(n-M+1)\right)^M-1}{{s_n}^{-d}c(n-M+1)-1}\\ &= O\left({s_n}^{-d}\left(n{s_n}^{-d}\right)^{M-1}\right)\end{aligned}$$

The second part of the statement follows directly from $|\cup_{m=1}^{M}\mathcal{N}^m(i)|\geq \mathcal{N}(i) > N$ for any N with probability approaching one.

Finally, for part b), the probability of interest is bounded from below as

$$\begin{aligned}\mathbb{P}[|\mathcal{N}_t(i;j)|\geq\tau] &\geq \mathbb{P}[|\mathcal{N}_t(i;j)|\geq\tau|X_j\in B_{s_n^{-1}}(x_i)]\mathbb{P}[X_j\in B_{s_n^{-1}}(x_i)]\\ &\geq \mathbb{P}[|\mathcal{N}_t(i;j)|\geq\tau|X_j\in B_{s_n^{-1}}(x_i)]C{s_n}^{-d}\end{aligned}$$

The number of friends in common, conditional on X_j being in the ball $B_{s_n^{-1}}(x_i)$, follows a binomial distribution with parameters
$(n-1, \mathbb{P}[W_{il}=W_{jl}=1|X_j\in B_{s_n^{-1}}(x_i)])$. The probability is bounded from below by $C{s_n}^{-d}$.

Indeed, noting that $\|x_j-x_k\|\leq s_n^{-1}+\|x_i-x_k\|$ when $x_j\in B_{s_n^{-1}}(x_i)$,

$$\begin{aligned}&\mathbb{P}[(1-w_{ik})(1-w_{jk})|X_j\in B_{s_n^{-1}}(x_i)]\\ &= \int_{\mathcal{X}}\int_{\mathcal{X}}(1-w_{ik})(1-w_{jk})\frac{f_{X_j}(x_j)}{\int_{B_{s_n^{-1}}(x_i)}}\frac{\mathbb{1}_{B_{s_n^{-1}}(x_i)}f_{X_j}}{\int_{B_{s_n^{-1}}(x_i)}f_{X_j}}f_{X_k}(x_k)\;dx_j\;dx_k\\ &\geq C{s_n}^{d}\int_{\mathcal{X}}\int_{B_{s_n^{-1}}(x_i)} g(s_n\|X_k-x_i\|)g(1+s_n\|X_k-x_i\|)f_{X_j}(x_j)f_{X_k}(x_k)\;dx_j\;dx_k\\ &\geq C{s_n}^{-d}\int_{\mathcal{X}} g(\|y\|)\;g(1+\|y\|)\;f_{X_k}\left(\frac{y}{s_n}+x_i\right)\;dx_k\\ &\geq C{s_n}^{-d}\;O(1)\end{aligned}$$

where the change of variable $s_n(x_k-x_i)\to y$ was used.

Since the Bernoulli trials from the binomial are conducted $n-1$ times with a probability at least of order ${s_n}^{-d}$, the probability of having at least τ friends in common approaches one as long as $n{s_n}^{-d}\to\infty$ as in a). Thus, $\mathbb{P}[|\mathcal{N}_t(i;j)|\geq\tau]\geq C\;\mathbb{P}[X_j\in B_{s_n^{-1}}(x_i)] = C{s_n}^{-d}$.

□

4.2 Theorem 2.1

Proof. **Part a)** Lemma 1 ensures asymptotics apply. By the overlap assumption, the number of treated friends and untreated friends both grow to infinity.

Consider the treated group. Letting $\tilde{Y}_{j;n}$ denote i.i.d. draws from $Y(1)|\mathcal{C}_{i1}$, we have $\sup_n \mathbb{E}[(\tilde{Y}_{j;n})^2] < \infty$ since $\mathbb{E}[Y_j(1)^2|\mathcal{C}_{i1}] \to \mathbb{E}[Y_j(1)^2|X_j = x_i, T_j = 1] = \mathbb{E}[Y_j(1)^2|X_j = x_i]$ by continuity (this is shown in details for $Y_j(1)$ below) and unconfoundedness. Then, $\frac{1}{\mathcal{C}_{i1}} \sum_{j \in \mathcal{C}_{i1}} (Y_j(T_j) - \mathbb{E}[Y_j(1)|\mathcal{C}_{i1}]) + \mathbb{E}[Y_j(1)|\mathcal{C}_{i1}] - \mathbb{E}[Y_j(1)|X_j = x_i] \to^p 0$ by the law of large numbers for triangular arrays and continuity, which is formally established by bounding $\mathbb{E}[Y_j(1)|\mathcal{C}_{i1}] - \mathbb{E}[Y_j(1)|X_j = x_i]$ as follows.

The expectation can be divided into an integral over a ε-ball centered at x_i and an integral over its complement. Within the ball,

$$\begin{aligned}
&\left| \int_{B_\varepsilon(x_i)} (\mathbb{E}[Y_j(1)|X = x] - \mathbb{E}[Y_j(1)|X = x_i]) f_{X|\mathcal{C}_i,T} \, dx \right| \\
&\leq \sup_{B_\varepsilon} |\mathbb{E}[Y_j(1)|X = x] - \mathbb{E}[Y_j(1)|X = x_i]| \\
&= O(\varepsilon^\alpha)
\end{aligned}$$

by Hölder continuity.

The integral outside the ball,
$\int_{B_\varepsilon^c(x_i)} (\mathbb{E}[Y_j(1)|X = x] - \mathbb{E}[Y_j(1)|X = x_i]) f_{X|\mathcal{C}_i,T} \, dx$, can be bounded using information about the composition of $\mathcal{C}_i$. I proceed with the two groups of counterfactuals separately.
For the first choice of counterfactuals, *i.e.*, $\mathcal{C}_i = \cup_{m=1}^M \mathcal{N}^m(i)$, note that $\varepsilon < \|x_j - x_i\| = \left\|\sum_{k=0}^{m-1} x_{j_{k+1}} - x_{j_k}\right\| \leq \sum_{k=0}^{m-1} \|x_{j_{k+1}} - x_{j_k}\|$ and thus outside the ball

$$\begin{aligned}
f_{\mathcal{C}_i|X_j,T_j} &= \sum_{m=1}^M \mathbb{P}[W_{ij}^m = 1, W_{ij}^{m'} = 0 \; m' < m|X, T] \\
&= \sum_{m=1}^M \mathbb{E}[\mathbb{P}[W_{ij}^m = 1, W_{ij}^{m'} = 0 \; m' < m|\{X_{j_k}\}, T]|X_j, T_j] \\
&\leq \sum_{m=1}^M \int \prod_{k=0}^{m-1} (1 - w_{j_{k+1}j_k}) \prod_{l \neq k} w_{j_k j_l} f_{X_{-j}|X_j} \\
&\leq \sum_{m=1}^M \left(1 - w_n \left(\frac{\varepsilon}{m}\right)\right)^m
\end{aligned}$$

It follows that

$$
\begin{aligned}
&\left|\int_{B_\varepsilon^c(x_i)} (\mathbb{E}[Y_j(1)|X=x] - \mathbb{E}[Y_j(1)|X=x_i]) f_{X|\mathcal{C}_i,T}\ dx\right| \\
&\leq \int_{B_\varepsilon^c(x_i)} |\mathbb{E}[Y_j(1)|X=x] - \mathbb{E}[Y_j(1)|X=x_i]|\ \frac{\mathbb{P}[\mathcal{C}_i|X=x] f_{X,T}(x,1)}{\mathbb{P}[\mathcal{C}_i, T=1]}\ dx \\
&\leq \frac{\sum_{m=1}^M \left(1 - w_n(\frac{\varepsilon}{m})\right)^m}{\left(\sum_{m=1}^M \sum_{j_0,\cdots,j_m} \mathbb{P}[W_{j_k j_{k+1}} = 1, W_{j_k j_l} = 0\ l \neq k+1]\right)} (\mathbb{E}[|\mathbb{E}[Y_j(1)|X]|] + \mathbb{E}[Y_j(1)|X=x_i]) \\
&\leq C \frac{n}{\lambda_n M} \sum_{m=1}^M \left(1 - w_n\left(\frac{\varepsilon}{m}\right)\right)^m
\end{aligned}
$$

Hence, the integrals are $O({\varepsilon_n}^\alpha)$ and $O(\frac{n}{\lambda_n M}\sum_{m=1}^M (1 - w_n(\frac{\varepsilon_n}{m}))^m)$, respectively, and the bias disappears provided $\sum_{m=1}^M (1 - w_n(\frac{\varepsilon_n}{m}))^m = o(\lambda_n \frac{M}{n})$. Applying the same reasoning to the non-treated gives the result for the first choice of group of counterfactuals.

For the second group of counterfactuals, the posterior (leaving the conditioning on $T = t$ implicit) reads

$$
\begin{aligned}
f_{X_j|\mathcal{C}_i} &= \frac{\mathbb{P}[\mathcal{C}_i|X_j] f_{X_j}(x)}{\int \mathbb{P}[\mathcal{C}_i|X_j] f_{X_j}(x)\ dx} \\
&= \frac{\mathbb{E}[\mathbb{P}[W_{i1} = W_{j1} = \cdots = W_{im} = W_{jm} = 1|X_1, \cdots, X_m, X_j]] f_{X_j}(x)}{\int \mathbb{E}[\mathbb{P}[W_{i1} = W_{j1} = \cdots = W_{im} = W_{jm} = 1|X_1, \cdots, X_m, X_j]] f_{X_j}(x)\ dx} \\
&= \frac{\prod_{\tau=1}^m \int (1 - w_n(\|x_i - x_k\|))(1 - w_n(\|x_j - x_k\|)) f_{X_k}(x_k) dx_k f_{X_j}(x)}{\prod_{\tau=1}^m \int (1 - w_n(\|x_i - x_k\|))(1 - w_n(\|x_j - x_k\|)) f_{X_k}(x_k) dx_k f_{X_j}(x)\ dx} \\
&\leq C {s_n}^d \sum_{m=\tau}^{n-2} \left(1 - w_n(\frac{\varepsilon_n}{2})\right)^m \\
&= C {s_n}^d \left(1 - w_n(\frac{\varepsilon_n}{2})\right)^\tau\ O(1)
\end{aligned}
$$

so that we want $\left(1 - w_n(\frac{\varepsilon_n}{2})\right)^\tau = o({s_n}^{-d})$

Part b) one can compute

$$\begin{aligned}
\mathbb{E}[Y_j(t)|\mathcal{C}_i, T_j = t] &= \mathbb{E}[\mathbb{E}[Y_j(t)|X_j]|\mathcal{C}_i, T_j = t] \\
&= \int_{\mathcal{X}} \mathbb{E}[Y_j(t)|x_j] f_{x_j|\mathcal{C}_i,x_i,t} dx_j \\
&= \int_{\mathcal{X}} \mathbb{E}[Y_j(t)|x_j] f_{x_j|\mathcal{C}_i,x_i,t} dx_j \\
&= \int_{\mathcal{X}} \mathbb{E}[Y_j(t)|x_j] \frac{(1-w_{ij}) p_t(x_j) f_{x_j}}{\int_{\mathcal{X}} (1-w_{ik}) p_t(x_k) f_{x_k} dx_k} dx_j \\
&= \int_{\mathcal{X}} \mathbb{E}[Y_j(t)|x_i + y/s_n] \frac{g(\|y\|) p_t(x_i + y/s_n) f_{x_i+y/s_n}}{\int_{\mathcal{X}} g(\|z\|) p_t(x_i - z/s_n) f_{x_i - z/s_n} dz} dy \\
&= \mathbb{E}[Y_j(t)|x_i] \int_{\mathcal{X}} \frac{g(\|y\|)}{\int_{\mathcal{X}} g(\|z\|) dz} dy + O({s_n}^{-1}) \int_{\mathcal{X}} g(\|y\|) y dy + O({s_n}^{-2}) \\
&= \mathbb{E}[Y_j(t)|x_i] + O({s_n}^{-2})
\end{aligned}$$

using the changes of variable $y = s_n(x_j - x_i)$ and $z = s_n(x_k - x_i)$ and a first-order expansion in densities, expectation, and propensity scores (noting that $\frac{A+a_n}{B+b_n} = \frac{A}{B}\left(1 + \frac{a_n}{A}\right)\left(1 - \frac{b_n}{B+b_n}\right) = \frac{A}{B} + O(a_n) + O(b_n) + O(a_n b_n)$).

□

4.3 Theorem 2.2

Proof. By Lindeberg-Feller's central limit theorem, where Lindeberg's condition follows from the dominated convergence theorem,

$$\begin{aligned}
&\sqrt{|\mathcal{C}_i|} \begin{pmatrix} \frac{1}{|\mathcal{C}_{i1}|} \sum_{j \in \mathcal{C}_{i1}} (Y_j(T_j) - \mathbb{E}[Y_j(1)|\mathcal{C}_i, T_j = 1]) \\ \frac{1}{|\mathcal{C}_{i0}|} \sum_{j \in \mathcal{C}_{i0}} (Y_j(T_j) - \mathbb{E}[Y_j(0)|\mathcal{C}_i, T_j = 0]) \end{pmatrix} \\
&\to^d \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} \frac{\mathbb{V}[Y_j(1)|X_j = x_i]}{\mathbb{P}[T_j=1|X_j=x_i]^2} & \frac{\mathbb{C}[Y_j(1);Y_j(0)|X_i=x_i]}{\mathbb{P}[T_j=1|X_j=x_i]\mathbb{P}[T_j=0|X_j=x_i]} \\ \frac{\mathbb{C}[Y_j(1);Y_j(0)|X_i=x_i]}{\mathbb{P}[T_j=1|X_j=x_i]\mathbb{P}[T_j=0|X_j=x_i]} & \frac{\mathbb{V}[Y_j(0)|X_j=x_i]}{\mathbb{P}[T_j=0|X_j=x_i]^2} \end{pmatrix} \right)
\end{aligned}$$

It is now required that $\sqrt{n}(\mathbb{E}[Y_j(t)|\mathcal{C}_i, T_j = t] - \mathbb{E}[Y_j(t)|X_j = x_i]) \to 0$ sufficiently fast. From the consistency proof, it is seen that the bias disappears faster than root n when $\varepsilon_n = n^{-\gamma}$ under the conditions of the theorem. □

4.4 Theorem 2.3

Consider $\frac{1}{n}\sum_{i=1}^n \hat{\omega}_i Y_i(t) = \frac{1}{n}\sum_{i=1}^n \omega_i Y_i(t) + \frac{1}{n}\sum_{i=1}^n (\hat{\omega}_i - \omega_i) Y_i(t)$. Here, the first term features the inverse propensity score weighting, ensuring the sample average converges to $\mathbb{E}[Y_i(t)]$, while the second term will be shown to be asymptotically negligible. Because of Markov inequality, it will suffice to show that $\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n (\hat{\omega}_i - \omega_i) Y_i(T_i)\right|\right] \leq$

$\mathbb{E}[\mathbb{E}[|(\hat{\omega}_i - \omega_i)Y_i(T_i)|]$ is of order lower than root n. Using Cauchy-Schwartz and the hypothesis that moments of Y_i exist, the result hinges on the order of $\mathbb{E}[(\hat{\omega}_i - \omega_i)^2|X_i, T_i, \mathcal{C}_i] = \mathbb{V}[\hat{\omega}_i|X_i, T_i, \mathcal{C}_i] + (\mathbb{E}[\hat{\omega}_i - \omega_i|X_i, T_i, \mathcal{C}_i])^2$. The bias term reads

$$\begin{aligned}
\mathbb{E}[(\hat{\omega}_i - \omega_i)|X_i, T_i, \mathcal{C}_i] &= \mathbb{E}\left[\frac{n\mathbb{P}[\mathcal{C}_i|X_i]\mathbb{1}_1(T_i)}{1+|\mathcal{C}_{j1}|} - \frac{\mathbb{1}_1(T_i)}{\mathbb{P}[T_i=1|X_i]}\middle| X_i, T_i, \mathcal{C}_i\right] \\
&\quad - \mathbb{E}\left[\frac{n\mathbb{P}[\mathcal{C}_i|X_i]\mathbb{1}_0(T_i)}{1+|\mathcal{C}_{j0}|} - \frac{\mathbb{1}_0(T_i)}{\mathbb{P}[T_i=0|X_i]}\middle| X_i, T_i, \mathcal{C}_i\right] \\
&= \frac{\mathbb{P}[\mathcal{C}_i|X_i]\mathbb{1}_1(T_i)}{\tilde{p}_1} - \mathbb{P}[\mathcal{C}_i|X_i]\mathbb{1}_1(T_i)\frac{(1-\tilde{p}_1)^n}{\tilde{p}_1} \\
&\quad - \frac{\mathbb{1}_1(T_i)}{\mathbb{P}[T_i=1|X_i]} \\
&\quad - \frac{\mathbb{P}[\mathcal{C}_i|X_i]\mathbb{1}_0(T_i)}{\tilde{p}_0} + \mathbb{P}[\mathcal{C}_i|X_i]\mathbb{1}_0(T_i)\frac{(1-\tilde{p}_0)^n}{\tilde{p}_0} \\
&\quad + \frac{\mathbb{1}_0(T_i)}{\mathbb{P}[T_i=0|X_i]}
\end{aligned}$$

where we used $\mathbb{E}\left[\frac{1}{1+X}\right] = \frac{1}{pn}(1-(1-p)^n)$ when X follows a $\mathcal{B}(n-1,p)$ distribution (Chao and Strawderman, 1972) and where (expressed here for $\mathcal{C}_i = \mathcal{N}_i$ for simplicity) $\tilde{p}_t \overset{\text{def}}{=} \int\int g(s_n\|x_k - x_j\|)f(x_k)\mathbb{P}[T=t|x_k]\frac{g(s_n\|x_j-x_i\|)f(x_j)}{\int g(s_n\|x_j-x_i\|)f(x_j)dx_j}dx_k dx_j$.

Recall $\mathbb{P}[\mathcal{C}_i|X_i] = {s_n}^{-d}\int g(\|y\|)f(x_i + \frac{y}{s_n})\ dy$ and change variables in $\tilde{p}_t$: $\tilde{p}_t = {s_n}^{-2d}\int\int g(\|y\|)f(x_i+\frac{y+z}{s_n})\mathbb{P}[T=t|x_j+\frac{y}{s_n}]\frac{g(\|z\|)f(x_i+z/s_n)}{\mathbb{P}[\mathcal{C}_i|X_i]}dzdy$. Then, similarly to the derivation of the CATE estimator's bias, a Taylor expansion for the propensity score and the density $f(x_i+y/s_n+z/s_n)$ yields $\frac{\mathbb{1}_t(T_i)}{\mathbb{P}[T_i=t|X_i]} + O({s_n}^{-2})$. The remaining terms obey $\mathbb{P}[\mathcal{C}_i|X_i]\mathbb{1}_t(T_i)(1-\tilde{p}_t)^n/\tilde{p}_t = O_p(e^{-\lambda_n})$ using $(1-\tilde{p}_t)^n = e^{n\ln(1-\tilde{p}_t)}$, developing the logarithm, and using previous rates.

The variance term can be analyzed similarly, noting $\mathbb{V}\left[\frac{1}{1+X}\right] = O((np)^{-2})$ when X follows a $\mathcal{B}(n-1,p)$ distribution (Cribari-Neto, Garcia and Vasconcellos, 2000).

Finally, note that the DCT applies. Indeed, $\mathbb{E}[1/\mathbb{P}[T_i = t|X_i]] \leq \mathbb{E}[1/C] = 1/C < \infty$ and $\mathbb{E}\left[\frac{\mathbb{P}[\mathcal{C}_i|X_i]}{\tilde{p}_t}\right] \leq 1/C < \infty$, where the latter follows from

$$\begin{aligned}
(\mathbb{P}[\mathcal{C}_i|X_i])^2 &= \left(\int_{\mathcal{X}} g(s_n\|x_j - x_i\|)f(x_j)\ dx_j\right)^2 \\
&= \left(\int_{\mathcal{X}} \frac{s_n^d g(s_n\|x_j - x_i\|)}{\int g(\|y\|)\ dy} f(x_j)\ dx_j\right)^2 \left({s_n}^{-d}\int g(\|y\|)\ dy\right)^2 \\
&= (\mathbb{E}_G[f(X_j)])^2 \\
&\leq \mathbb{E}_G[f(X_j)^2] \\
&= \int_{\mathcal{X}} g(s_n\|x_j - x_i\|)f(x_j)^2\ dx_j {s_n}^{-d}\left(\int g(\|y\|)\ dy\right)^2
\end{aligned}$$

by Jensen's inequality, noting that $\frac{s_n^d g(s_n\|x_j - x_i\|)}{\int g(\|y\|)\ dy}$ constitutes a density G.

Furthermore,

$$\begin{aligned}
\mathbb{P}[\mathcal{C}_i|X_i]\tilde{p}_t &= \int\int g(s_n\|x_k - x_j\|)\ f(x_k)\mathbb{P}[T=t|x_k]\ g(s_n\|x_j - x_i\|)\ f(x_j)\ dx_k\ dx_j \\
&\geq C{s_n}^{-d}\int_{\mathcal{X}} f(x_j)\left(\int_{\mathcal{X}_y} g(\|y\|)f(x_j + y/s_n)\ dy\right)\ g(s_n\|x_j - x_i\|)\ dx_j \\
&\geq C{s_n}^{-d}\int f(x_j)^2\ g(s_n\|x_j - x_i\|)\ dx_j
\end{aligned}$$

noting $\int_{\mathcal{X}_y} g(\|y\|)f(x_j + y/s_n)\ dy \geq g(\varepsilon)\int_{\mathcal{X}_y \cap B_\varepsilon(0)} f(x_j + y/s_n) \geq g(\varepsilon)V(\mathcal{X}_y \cap B_\varepsilon(0))\inf_{\tilde{x}\in\mathcal{X}} \frac{\int_{B_\varepsilon(0)} f(\tilde{x}+y/s_n)\ dy}{f(\tilde{x})} f(x_j)$ and the infimum is bounded away from 0 uniformly over n as (i) $\int_B f \geq f$ on K^c by convexity of tails[13] and (ii) if the infimum is reached on K, then by compactness there exists a subsequence that converges to x^*, the infimum is reached, and a value of 0 would imply $B_\varepsilon(x^*) \subset supp(f)^c$. As the ratio converges to 1 as $s_n \to \infty$, it can be concluded that $\inf_{\tilde{x}\in\mathcal{X}} \frac{\int f(\tilde{x}+y/)\ dy}{f(\tilde{x})} \geq C > 0$.

As a result, $\mathbb{E}\left[\frac{\mathbb{P}[\mathcal{C}_i|X_i]}{\tilde{p}_t}\right] \leq C$.

4.5 Theorem 2.4

Proof. To ease notation, the argument κ is omitted in all instances of $\mathcal{C}_{it}(\kappa)$. I also make use of the following shorthands: $\Delta_{ij}^u \overset{\text{def}}{=} \|x_j^u - x_i^u\|$, $h_{ij} \overset{\text{def}}{=} h(X_j^o, X_i^o)$.

Decompose the centered mean of the group as

$$\frac{1}{|\mathcal{C}_{it}|}\sum_{j\in\mathcal{C}_{it}} Y_j(T_j) - \mathbb{E}[Y(t)|x_i] = \frac{1}{|\mathcal{C}_{it}|}\sum_{j\in\mathcal{C}_{it}} Y_j(T_j) - \mathbb{E}[Y_j(t)|\mathcal{C}_{it}] + \mathbb{E}[Y_j(t)|\mathcal{C}_{it}] - \mathbb{E}[Y(t)|x_i]$$

for $t \in \{0, 1\}$.

The first term depends on sample fluctuations and converges (in probability to 0 and in distribution once re-scaled) provided the number of observations in the sum grows to infinity. The second term is a bias term and disappears under regularity conditions and the truncation $h_{ij} > \kappa$ with $\kappa \to \infty$.

Thus, the main steps prove that (i) the number of observations in the sum grows to infinity and (ii) the bias disappears, for some sequence $\kappa \to \infty$. Consider the

[13] This is immediate by Taylor expansion if the density is twice continuously differentiable. The result can be established by a mollification argument in general.

probability of an observation belonging to $\mathcal{C}_i$ first:

$$\begin{aligned}
\mathbb{P}[\mathcal{C}_i] &\geq C \int_{h_{ij}>\kappa} (1 - w_{ij}) f_{X_j}(x_j)\ dx_j \\
&\geq C \int_{\Delta_{ij}^u \leq \frac{b_n}{2}, \kappa < h_{ij} < \kappa + \frac{b_n}{2}} (1 - w(\kappa + b_n)) f_{X_j}(x_j)\ dx_j \\
&= C(1 - w(\kappa + b_n))\mathbb{P}\left[\Delta_{ij}^u \leq \frac{b_n}{2}, \kappa < h_{ij} < \kappa + \frac{b_n}{2}\right] \\
&= C(1 - w(\kappa + b_n))\mathbb{P}\left[\Delta_{ij}^u \leq \frac{b_n}{2}\right] \mathbb{P}\left[\kappa < h_{ij} < \kappa + \frac{b_n}{2} \middle| \Delta_{ij}^u \leq \frac{b_n}{2}\right] \\
&\geq C(1 - w(\kappa + b_n)){b_n}^{d_u+1}
\end{aligned}$$

Following previous proofs in the *strong homophily* case, it suffices to let $\kappa \to \infty$ slowly enough as to induce a rate of $\frac{\lambda_n}{n}$ for the above probability.

Next, let's turn to the bias term. It reads

$$\begin{aligned}
|\mathbb{E}[Y_j(T_j)|\mathcal{C}_{it}] - \mathbb{E}[Y_j(t)|x_i]| &= ||\mathbb{E}[\mathbb{E}[Y_j(T_j)|X_j]|\mathcal{C}_{it}]] - \mathbb{E}[Y_j(t)|x_i]|| \\
&= \left| \int_{\mathcal{X}} (\mathbb{E}[Y_j(t)|x_j] - \mathbb{E}[Y_j(t)|x_i]) f_{X_j|\mathcal{C}_{it}}(x_j)\ dx_j \right| \\
&\leq \left| \int_{B_\varepsilon(x_i)} (\mathbb{E}[Y_j(t)|x_j] - \mathbb{E}[Y_j(t)|x_i]) f_{X_j|\mathcal{C}_{it}}(x_j)\ dx_j \right| \\
&+ \left| \int_{B_\varepsilon^c(x_i)} (\mathbb{E}[Y_j(t)|x_j] - \mathbb{E}[Y_j(t)|x_i]) f_{X_j|\mathcal{C}_{it}}(x_j)\ dx_j \right| \\
&\leq C\varepsilon^\alpha \\
&+ \frac{1}{\mathbb{P}[\mathcal{C}_{it}]} \int_{B_\varepsilon^c(x_i)} (\mathbb{E}[Y_j(t)|x_j] - \mathbb{E}[Y_j(t)|x_i]) \mathbb{P}[\mathcal{C}_{it}|x_j] f_{X_j,T_j}(x_j, t)\ dx_j \\
&\leq C\varepsilon^\alpha \\
&+ \frac{C}{\mathbb{P}[\mathcal{C}_{it}]} \int_{B_\varepsilon^c(x_i), h_{ij}>\kappa} (\mathbb{E}[Y_j(t)|x_j] - \mathbb{E}[Y_j(t)|x_i])(1 - w_{ij}) f_{X_j,T_j}(x_j, t)\ dx_j \\
&\leq C\varepsilon^\alpha + \frac{Cn}{\lambda_n} (1 - w(\kappa + \varepsilon) (\mathbb{E}[|\mathbb{E}[Y_j(t)|X_j]||T_j = t] + |\mathbb{E}[Y_j(t)|x_i])
\end{aligned}$$

so that, with $\varepsilon \downarrow 0$, the bias disappears as κ rises provided $(1 - w(\kappa + \varepsilon)) = o(\frac{\lambda_n}{n})$. □

4.6 Theorem 2.5

Proof. The proof proceeds along the same line as that of the previous theorem, going through the same main two steps. The only changes come from modifying the

bounds on the probability that an individual belongs to the group of counterfactuals and on that probability conditional on X_j.

Noting that $\mathbb{P}[\mathcal{C}_{it}|X_j] = n\mathbb{E}[(1-w_{ik})(1-w_{ikjk})\mathbb{1}_{h_{ik}>\kappa,h_{jk}>\kappa}|X_j]$, the probability that an individual belongs to the group of counterfactuals can be bounded as follows:

$$\begin{aligned}\mathbb{P}[\mathcal{C}_{it}] &\geq Cn(1-w(\kappa+b_n))^2\ \mathbb{P}\left[h_{ik}\in\left[\kappa,\kappa+\frac{b_n}{2}\right],h_{jk}\in\left[\kappa,\kappa+\frac{b_n}{2}\right],\Delta^u_{ik}\leq\frac{b_n}{2},\Delta^u_{jk}\leq\frac{b_n}{2}\right]\\ &\geq Cn\left((1-w(\kappa+b_n))b_n{}^{d_u+1}\rho_h\left(\kappa+\frac{b_n}{2}\right)\right)^2\end{aligned}$$

where we used

$$\begin{aligned}\mathbb{P}\left[\Delta^u_{ik}\leq\frac{b_n}{2},\Delta^u_{jk}\leq\frac{b_n}{2}\right] &= \mathbb{P}\left[\Delta^u_{ik}\leq\frac{b_n}{2}\right]\mathbb{P}\left[\Delta^u_{jk}\leq\frac{b_n}{2}\middle|\Delta^u_{ik}\leq\frac{b_n}{2}\right]\\ &\geq C(b_n/2)^{d_u}(b_n/2)^{d_u}\end{aligned}$$

and a similar reasoning for the term involving h, under the homophily assumption.

Finally, on $B^c_\varepsilon(x^u_i)$, we get an upper bound for the probability that an individual belongs to the group of counterfactuals conditional on X_j from
$\mathbb{P}[\mathcal{C}_{it}|X_j] = n\mathbb{E}[(1-w_{ik})(1-w_{ikjk})\mathbb{1}_{h_{ik}>\kappa,h_{jk}>\kappa}|X_j] \leq Cn(1-w(\kappa+\frac{\varepsilon}{2}))(1-w(\kappa))$
□

Appendix B: Parametric models and Extensions

The insight that motivated the estimators analyzed in the main text lends itself to various formalizations. Besides variations on the network formation process or additional parametric restrictions, one could replace truncation-based estimators with a K-nearest neighbors version, for instance by selecting the K^o closest individuals to i in terms of observables and then using the $K < K^o$ people with the largest values of h among the friends of i to sum over. In general, it would be expected to deliver similar results, possibly with the usual kernel vs. nearest-neighbors differences. In addition to varying the preferred nonparametric method, however, one can also apply the main method to more parametrized models, *e.g.*, logit propensity score or regression, to simplify the analysis, construct doubly-robust estimators, and possibly make use of more observations. I illustrate the idea in the propensity score and regression cases.

4.7 Propensity score analysis with maximum likelihood

Suppose the propensity score depends on both observables and unobservables and can be parametrized as $\mathbb{P}[T_j = 1|X_j] = G(\alpha + X_j\beta)$.

I consider a maximum likelihood estimator which maximizes $\frac{1}{|\mathcal{N}(i;\kappa)|}\sum_{j\in\mathcal{N}(i;\kappa)} f_j$ where $f_j \overset{\text{def}}{=} T_j \ln(G(\alpha + X_j\beta) + (1 - T_j)\ln(1 - G(\alpha + X_j\beta))$. Under usual extremum estimator assumptions (*i.e.*, compactness: $(\alpha, \beta) \in \mathcal{A} \times \mathcal{B}$ is compact, dominance: $f_j(x^o, x^u, \beta) \leq d(x^o, x^u)$ is integrable), I show how the previous manipulations deliver consistent estimator absent data on X^u, which is assumed to have bounded support for ease of argument. I focus on the κ-truncation case here, but the reasoning extends to variations such as the *strong homophily* cases.

The main difficulty in applying standard consistency argument for maximum of likelihood or extremum estimator in general (Newey and McFadden, 1994) lie in showing uniform convergence in probability, which is argued here.

Note that $\tilde{\alpha}_i \overset{\text{def}}{=} \alpha + X_j^u\beta_u$ converges, as the neighborhood shrink upon increasingly similar friends in terms of unobservables, to $\alpha + x_i^u\beta_u$. Consider now

$\sup_{\tilde{\alpha},\beta} \frac{1}{|\mathcal{C}_i(\kappa)|}\sum_{j\in\mathcal{C}_i(\kappa)} f_j - \mathbb{E}[f_j|X_j^u = x_i^u]$

$= \sup_{\tilde{\alpha},\beta} \left(\frac{1}{|\mathcal{N}_i(\kappa)|}\sum_{j\in\mathcal{C}_i(\kappa)} f_j - \mathbb{E}[f_j|\mathcal{C}_i, h \geq \kappa^u] + \mathbb{E}[f_j|\mathcal{C}_i, h \geq \kappa] - \mathbb{E}[f_j|X_j^u = x_i^u] \right)$

This can be broken down into two suprema. The first converges by a uniform law of large numbers (compactness follows by assumption; the dominating function satisfies $\mathbb{E}[d(x)|\mathcal{C}_i, h \geq \kappa] \leq \int_{B_\varepsilon} d(x)dx + Cn^\beta(1 - w(\kappa+\varepsilon)) \int_{B_\varepsilon^c} d(x)f(x)dx < \infty$, and thus is integrable uniformly over the conditioning sets). The second supremum can be bounded as in the previous section to obtain a bound whose only dependence on (α, β) is the (Hölder) modulus of continuity so that compactness delivers a uniform bound which shrinks to 0.

Hence we have the following result: for any i, the MLE based on $\mathcal{N}(i;\kappa)$, $\hat{\beta}_i$, satisfies $\hat{\beta}_i \to^p \beta$ under the usual binary outcome compactness and dominance assumptions, supplemented by the network formation conditions of previous theorems.

The resulting estimators of β (based on each i- neighborhood) can be improved by averaging over i, under optimal weighting. Note that the correlation between the estimators disappears as neighborhood of friends with close unobservables eventually stop overlapping.

Combined with consistent estimation of the $\tilde{\alpha}_i$, it is then possible to fit a propensity score for each individual from which treatment effect estimates follow.

4.8 Regression

Consider

$$y = \alpha + X\beta + \tau T + \varepsilon$$

For given i, y can be regressed on X^o, T using observations that constitute good counterfactuals for i. By similar arguments to before, the estimates will converge to (β_o, CATE) as the sample size rises. Since the model imposes more structure however, better estimators are available. In particular, treatment effects are homogeneous and a more efficient (C)ATE estimator is formed by averaging over i.

Put observed regressors and treatment status into the vector a_j. Using Frisch-Waugh and denoting expectation-centering by a double dot, $\begin{pmatrix} \beta_o^{\mathcal{N}(i)} \\ \tau^{\mathcal{N}_i} \end{pmatrix} = \begin{pmatrix} \beta_o \\ \tau \end{pmatrix} + (\mathbb{E}[\ddot{a}_j \ddot{a}_j' | \mathcal{C}_i])^{-1} (\mathbb{E}[\ddot{a}_j (x_j^u)' | \mathcal{C}_i]) \beta_u$

The linear model translates to a more tractable form for the bias, with the (inverse of) $\mathbb{E}[\ddot{a}_j \ddot{a}_j' | \mathcal{C}_i]$ being estimable. The term $\mathbb{E}[\ddot{a}_j (x_j^u)' | \mathcal{N}_i] = \int \ddot{a}_j (x_j^u)' f_{X|\mathcal{C}_i} \, dx$ can be analyzed as the bias has been previously, but the particular form of the integral allows for some tractable bounding under some assumptions.

(i) using Hölder, $\leq \mathbb{E}[\|\ddot{a}_j\|^2 | E_w] \mathbb{E}[\|x_j^u\|^2 | \mathcal{C}_i]$

(ii) If $w = 1_{\|x\| \leq K}$, then $\leq (\|x_i^u\| + K) \mathbb{E}[\|\ddot{a}_j\| | \mathcal{C}_i]$.

Noting expectations involving a_j are identified, the bias can be bounded under restrictions on the effect of unobservables (*e.g.*, β_u lives in some compact set) and possibly bounding moments of unobservables (second moment in (i)). Of course, since the more interesting estimator averages out over i, the relevant bias is rather the expectation of the above term, which can be bounded similarly.

Finally, note the intercept for each sub-regression: it will asymptotically estimate $\alpha + (x_i^u)' \beta_u$ and thus allow to correct for unobservables at the individual level.

4.9 Common interest formation process

Many types of network formation have been investigated in the literature and homophilic behavior has different implications depending on the linking model. In general, homophily allows us to extract some information to refine counterfactuals, albeit with different quality and asymptotic behavior.

For another example that still uses dyadic link formation, consider a common interest formation process: i an j are friends if $|X_{ik} - X_{jk}| < D_k$ (or possibly η_{ijk} instead of D) for some k, where k ranges over the covariate indices. In this case, two persons become friends if they have some kind of common interest or have some characteristic in common.

This model rationalizes the idea that, if two persons have apparently nothing in common, it is likely they share some unobserved common interest, ability, or taste. If only one X_{ik} is unobserved then two individuals differing sufficiently on observables must be great counterfactuals for each other in terms of this X_{ik}. Such link formation processes have different implications for CATE estimation than the earlier link formation model. With univariate unobservables and $D = D_n = o(n^{-\beta}), \beta < 1$ a simple consistent estimator follows readily, while it is not possible to handle multivariate unobservables and truncation based on observables is helpless.

4.10 Heterogeneity and Popularity

One feature of empirical networks is heterogeneity in degree, *i.e.*, the number of connections varies widely from one individual to another. The frameworks presented in the main text provide two sources of such heterogeneity: (i) for given characteristics x_i, the noise η_{ij} provides some variation across individuals with similar attributes, and (ii) variations in characteristics induce various probabilities of forming connections; individuals who have more common characteristics have an easier term forming connections.

Thus, the models accommodate degree heterogeneity without further modification. It may however be desirable to incorporate additional variables that reflect "popularity" or "expansiveness" in some applications. This is compatible with the estimators developed in the main text. I briefly discuss two ways of doing so; a detailed analysis of the consequences of introducing popularity in different ways is beyond the scope of the paper.

One might introduce scale variables S_i to represent popularity as in $W_{ij} = \mathbb{1}_{S_i w(h(X_i^o; X_j^o) + \|X_i^u - X_j^u\|) + (1 - S_i) \leq \eta_{ij}}$. In case of undirected links, one may replace S_i by $\frac{S_i + S_j}{2}$ in the above expression to preserve symmetry. This would account for extraversion or popularity-type of characteristics in link formation by inducing heterogeneity in the probability of friendship for a given distance in the X variables.

These modifications would not alter the results under suitable regularity conditions (the most trivial imposing $0 < C < S_i < C < 1$ almost surely to induce straightforward bounds, but this can be relaxed).

Anoter possible way to incorporate popularity is by adding a term $(2 - P_i)(2 - P_j)\|P_i - P_j\|$ to $h_{ij} + \Delta^u_{ij}$ where $P_i \in [0; 1]$ is a popularity score. This specification reflects two facts about popularity: (i) popular people tend to have friends and (ii) popular people tend to be friends with popular people.

4.11 A note on directed and weighted links

It is generally possible to use the forthcoming estimators in the case of directed or weighted links. The former would no longer assume $\eta_{ij} = \eta_{ji}$, while the latter could use a model such as $W_{ij} = (w_n(h(X^o_i, X^o_j) + \|X^u_i - X^u_j\|) + \eta_{ij})/2$. Since the information contained in the network is richer in the cases, one can likely obtain better estimators, further information on the distribution of unobservables, or specification tests. An analysis of these cases is beyond the scope of this paper but would constitute a valuable and interesting exercise.

Appendix C: Simulation Results

Table 4: Performance (in terms of RMSE for ATE) of estimators

			M					τ			M, $\tau = 1$	OLS	Kernel
β_3	y	w	1	2	3	4	5	1	2	3			
0	A	a	0.24	0.15	0.13	0.13	0.13	0.15	0.21	0.30	0.24	0.15	0.21
		b	0.21	0.13	0.12	0.13	0.16	0.14	0.26	0.36	0.28		
		c	0.21	0.12	0.13	0.17	0.22	0.13	0.28	0.41	0.32		
		d	0.19	0.12	0.15	0.20	0.26	0.12	0.27	0.41	0.31		
	B	a	0.26	0.17	0.14	0.13	0.13	0.17	0.23	0.32	0.26	0.18	0.27
		b	0.22	0.14	0.12	0.13	0.16	0.15	0.27	0.37	0.28		
		c	0.23	0.13	0.13	0.17	0.21	0.14	0.29	0.43	0.34		
		d	0.23	0.13	0.14	0.17	0.22	0.14	0.31	0.45	0.36		
0.5	A	a	0.19	0.13	0.12	0.14	0.17	0.13	0.16	0.24	0.19	0.37	0.48
		b	0.16	0.12	0.15	0.19	0.23	0.12	0.20	0.29	0.21		
		c	0.16	0.12	0.18	0.24	0.31	0.12	0.22	0.35	0.26		
		d	0.15	0.13	0.20	0.27	0.34	0.12	0.23	0.37	0.27		
	B	a	0.16	0.12	0.12	0.15	0.18	0.12	0.14	0.21	0.16	0.48	0.62
		b	0.14	0.12	0.15	0.20	0.25	0.12	0.17	0.27	0.19		
		c	0.14	0.14	0.20	0.27	0.34	0.13	0.19	0.31	0.23		
		d	0.14	0.15	0.22	0.31	0.38	0.14	0.20	0.33	0.24		
1	A	a	0.18	0.12	0.13	0.16	0.20	0.12	0.15	0.23	0.18	0.63	0.70
		b	0.16	0.13	0.17	0.22	0.27	0.13	0.19	0.28	0.21		
		c	0.15	0.14	0.21	0.29	0.36	0.13	0.20	0.33	0.24		
		d	0.15	0.16	0.24	0.33	0.41	0.14	0.21	0.34	0.26		
	B	a	0.12	0.13	0.18	0.23	0.28	0.13	0.12	0.15	0.12	0.86	0.94
		b	0.12	0.15	0.21	0.28	0.35	0.14	0.13	0.20	0.14		
		c	0.12	0.19	0.28	0.37	0.46	0.17	0.14	0.24	0.17		
		d	0.12	0.20	0.30	0.40	0.50	0.18	0.15	0.27	0.19		

RMSE of the estimator using friends up to order M, the estimator using people with at least τ friends in common, the estimator using friends with at least another friend in common, OLS, and kernel regression.

Table 5: Performance (in terms of RMSE for CATE) of estimators

			M					τ			M, $\tau = 1$	OLS	Kernel
β_3	y	w	1	2	3	4	5	1	2	3			
0	A	a	0.85	0.57	0.45	0.41	0.41	0.57	0.71	0.75	0.85	0.15	0.25
		b	0.80	0.53	0.44	0.42	0.42	0.53	0.62	0.69	0.78		
		c	0.83	0.49	0.41	0.41	0.45	0.51	0.63	0.81	0.82		
		d	0.86	0.51	0.42	0.43	0.48	0.52	0.72	0.88	0.86		
	B	a	0.86	0.60	0.51	0.48	0.49	0.60	0.72	0.76	0.86	0.70	0.74
		b	0.81	0.53	0.47	0.45	0.48	0.54	0.63	0.71	0.81		
		c	0.89	0.55	0.47	0.50	0.57	0.56	0.70	0.85	0.88		
		d	0.88	0.58	0.49	0.53	0.61	0.59	0.75	0.90	0.89		
0.5	A	a	0.85	0.57	0.47	0.43	0.45	0.57	0.69	0.76	0.85	0.40	0.49
		b	0.79	0.53	0.45	0.44	0.48	0.53	0.62	0.70	0.77		
		c	0.84	0.50	0.44	0.48	0.54	0.52	0.67	0.82	0.84		
		d	0.86	0.53	0.46	0.50	0.57	0.55	0.71	0.89	0.88		
	B	a	0.88	0.62	0.53	0.53	0.56	0.62	0.73	0.77	0.88	0.84	0.94
		b	0.85	0.59	0.51	0.52	0.59	0.59	0.65	0.73	0.81		
		c	0.87	0.57	0.52	0.59	0.70	0.59	0.72	0.86	0.87		
		d	0.90	0.57	0.54	0.63	0.75	0.59	0.75	0.93	0.92		
1	A	a	0.85	0.59	0.48	0.46	0.49	0.59	0.71	0.77	0.85	0.67	0.71
		b	0.80	0.57	0.49	0.49	0.54	0.56	0.63	0.72	0.79		
		c	0.83	0.52	0.48	0.55	0.62	0.54	0.68	0.81	0.85		
		d	0.84	0.54	0.51	0.58	0.66	0.55	0.70	0.85	0.87		
	B	a	0.88	0.64	0.58	0.59	0.64	0.64	0.75	0.78	0.88	1.09	1.19
		b	0.83	0.60	0.54	0.58	0.68	0.61	0.67	0.75	0.82		
		c	0.89	0.61	0.60	0.71	0.85	0.63	0.73	0.86	0.89		
		d	0.91	0.61	0.63	0.75	0.88	0.62	0.76	0.93	0.90		

RMSE of the estimator using friends up to order M, the estimator using people with at least τ friends in common, the estimator using friends with at least another friend in common, OLS, and kernel regression.