The network interference model for treatment effect estimation places experimental units at the vertices of an undirected exposure graph, such that treatment assigned to one unit may affect the outcome of another unit if and only if these two units are connected by an edge. This model has recently gained popularity as means of incorporating interference effects into the Neyman–Rubin potential outcomes framework; and several authors have considered estimation of various causal targets, including the direct and indirect effects of treatment. In this paper, we consider large-sample asymptotics for treatment effect estimation under network interference in a setting where the exposure graph is a random draw from a graphon. When targeting the direct effect, we establish a central limit theorem and find that—in our setting—popular estimators are considerably more accurate than existing results suggest. Meanwhile, when targeting the indirect effect, we leverage our generative assumptions to propose a consistent estimator in a setting where no other consistent estimators are currently available. Overall, our results highlight the promise of random graph asymptotics in understanding the practicality and limits of causal inference under network interference.

1. Introduction. In many application areas, we seek to estimate causal effects in the presence of cross-unit interference, i.e., when treatment assigned to one unit may affect observed outcomes for other units. One popular approach to modeling interference is via an exposure graph or network, where units are placed along vertices of a graph and any two units are connected by an edge if treating one unit may affect exposure of the other: For example, Athey, Eckles and Imbens (2018) and Leung (2020) discuss experiments whose study units may interact via a social network, e.g., a friendship or professional network, and consider network interference models whose exposure graph corresponds to this social network. The statistical challenge is then to identify and estimate causal quantities in a way that is robust to such interference.

The existing literature on treatment effect estimation under network interference is formalized using a generalization of the strict randomization inference approach introduced by Neyman (1923). In a sense made precise below, these papers take both the interference graph and a set of relevant potential outcomes as deterministic, and then consider inference that is entirely driven by random treatment assignment (Aronow and Samii, 2017; Hudgens and Halloran, 2008). A major strength of this approach is that any conclusions derived from it are simple to interpret because they do not rely on any stochastic assumptions on either the outcomes or the interference graph. However, despite the transparency of the resulting analyses, it is natural to ask about the cost of using such strict randomization inference. If an approach to inference needs to work uniformly for any possible set of potential outcomes and any interference graph, does this limit its power over “typical” problems? Can appropriate stochastic assumptions enable more tractable analyses of treatment effect estimation under network interference, thus pointing the way to useful methodological innovations?
In this paper, we investigate the problem of treatment effect estimation under random graph asymptotics; specifically, we assume that the interference graph is a random draw from an (unknown) graphon. When paired with a number of regularity assumptions discussed further below, including an anonymous interference assumption, we find that our use of such random graph asymptotics lets us obtain considerably stronger guarantees than are currently available via randomization inference. When estimating direct effects, we find that standard estimators used in the literature are unbiased and asymptotically Gaussian for substantially denser interference graphs than was known before. And, when estimating indirect effects, our analysis guides us to a new estimator that has non-negligible power in a setting where no existing results based on randomization inference are available.

1.1. Graphon Asymptotics for Network Interference. Suppose that we collect data on subjects indexed \( i = 1, \ldots, n \), where each subject is randomly assigned a binary treatment \( W_i \sim \text{Bernoulli}(\pi) \) for some \( 0 \leq \pi \leq 1 \), and then experiences an outcome \( Y_i \). Following the Neyman-Rubin causal model (Imbens and Rubin, 2015), we posit the existence of potential outcomes \( Y_i(w) \in \mathbb{R} \) for all \( w \in \{0, 1\}^n \), such that the observed outcomes satisfy \( Y_i = Y_i(W) \). For notational convenience, we will often write \( Y_i(w_j = x; W_{-j}) \) to reference specific potential outcomes; here, \( Y_i(w_j = x; W_{-j}) \) means the outcome we would observe for the \( i \)-th unit if we assigned the \( j \)-th unit to treatment status \( x \in \{0, 1\} \), and otherwise maintained all but the \( j \)-th unit at their realized treatments \( W_{-j} \in \{0, 1\}^{n-1} \). We sometimes use shorthand \( Y_i(x; W_{-i}) := Y_i(w_i = x; W_{-i}) \) for the \( i \)-th index. Finally, we posit a graph with edge set \( \{E_{ij}\}_{i,j=1}^n \) and vertices at the \( n \) experimental subjects that constrains how potential outcomes may vary with \( w \): The \( i \)-th outcome may only depend on the \( j \)-th treatment assignment if there is an edge from \( i \) to \( j \), i.e., \( Y_i(w) = Y_i(w') \) if \( w_i = w'_i \) and \( w_j = w'_j \) for all \( j \neq i \) with \( E_{ij} = 1 \).

We seek to estimate the direct, indirect and total effects of the treatment on the outcome,

\[
\begin{align*}
\bar{\tau}_{\text{DIR}}(\pi) &= \frac{1}{n} \sum_{i} \mathbb{E}_\pi \left[ Y_i(w_i = 1; W_{-i}) - Y_i(w_i = 0; W_{-i}) \right], \\
\bar{\tau}_{\text{IND}}(\pi) &= \frac{1}{n} \sum_{i} \sum_{j \neq i} \mathbb{E}_\pi \left[ Y_j(w_i = 1; W_{-i}) - Y_j(w_i = 0; W_{-i}) \right], \\
\bar{\tau}_{\text{TOT}}(\pi) &= \frac{d}{d\pi} \left\{ \frac{1}{n} \sum_{i} \mathbb{E}_\pi [Y_i] \right\},
\end{align*}
\]

where the expectations above are taken over the random treatment assignment \( W_i \sim \text{Bernoulli}(\pi) \). This definition \( \bar{\tau}_{\text{DIR}} \) of the direct effect is by now standard (Halloran and Struchiner, 1995; Sävje, Aronow and Hudgens, 2021), while \( \bar{\tau}_{\text{IND}} \) is a formal analogue of this definition for the indirect effect. These estimands are further discussed by Hu, Li and Wager (2021), who show that in any Bernoulli experiment (and including in our current setting), \( \bar{\tau}_{\text{DIR}} \) and \( \bar{\tau}_{\text{IND}} \) decompose the total effect \( \bar{\tau}_{\text{TOT}} \), i.e., \( \bar{\tau}_{\text{TOT}}(\pi) = \bar{\tau}_{\text{DIR}}(\pi) + \bar{\tau}_{\text{IND}}(\pi) \).

Given a sampling model on the potential outcomes, we also consider limiting population estimands

\[
\tau_{\text{TOT}}(\pi) = \lim_{n \to \infty} \mathbb{E} \left[ \bar{\tau}_{\text{TOT}}(\pi) \right], \quad \tau_{\text{DIR}}(\pi) = \lim_{n \to \infty} \mathbb{E} \left[ \bar{\tau}_{\text{DIR}}(\pi) \right], \quad \ldots
\]

\[1\] Several recent papers have also considered network interference in completely randomized experiments where the number of treated units is fixed (e.g., in our setting, \( n_1 = \lfloor n \pi \rfloor \) randomly chosen units are assigned to treatment). This, however, gives rise to a number of subtle difficulties when studying estimands of the type (1) because treatment assignment across different units is not independent and so, in general, \( \mathbb{E}_\pi \left[ Y_j(w_i = x; W_{-i}) \neq \mathbb{E}_\pi \left[ Y_j(w_i = x; W_{-i}) \right] \mid W_i = x \right] \); see Sävje, Aronow and Hudgens (2021) and VanderWeele and Tchetgen (2011) for further discussion. Throughout this paper, we avoid such issues by only considering Bernoulli-randomized experiments.
provided these limiting objects exist. In this paper, we focus on estimating the quantities \( \tau_{\text{DIR}}(\pi), \tau_{\text{IND}}(\pi), \tau_{\text{TOT}}(\pi) \), etc, at the treatment probability \( \pi \) used for data collection.

Qualitatively, the total effect captures the effect of an overall shift in treatment intensity, while the direct effect captures the marginal responsiveness of a subject to their own treatment. Notice that the classical no-interference setting where \( Y_i \) only depends on the treatment assigned to the \( i \)-th unit is a special case of this setting with a null edge set; moreover, in the case without interference, \( \tau_{\text{TOT}}(\pi) \) and \( \tau_{\text{DIR}}(\pi) \) match and are equal to the sample average treatment effect, while the indirect effect is 0.

In the existing literature on treatment effect estimation under network interference, both the potential outcomes \( Y_i(w) \) and the edge set \( E_{ij} \) are taken as deterministic, and inference is entirely driven by the random treatment assignment \( W_i \sim \text{Bernoulli}(\pi) \) (Aronow and Samii, 2017; Athey, Eckles and Imbens, 2018; Basse, Feller and Toulis, 2019; Leung, 2020; Sävje, Aronow and Hudgens, 2021). This strict randomization-based approach, however, may limit the power with which we can estimate the causal quantities (1), and judicious stochastic modeling may help guide methodological advances in causal inference under interference. To this end, we consider a variant of the above setting that makes the following additional assumptions:

**Assumption 1 (Undirected Relationships).** The interference graph is undirected, i.e., \( E_{ij} = E_{ji} \) for all \( i \neq j \).

**Assumption 2 (Random Graph).** The interference graph is randomly generated as follows. Each subject has a random type \( U_i \sim \text{Uniform}[0, 1] \), and there is a symmetric measurable function \( G_n : [0, 1]^2 \to [0, 1] \) called a graphon such that \( E_{ij} \sim \text{Bernoulli}(G_n(U_i, U_j)) \) independently for all \( i < j \).

**Assumption 3 (Anonymous Interference).** The potential outcomes do not depend on the identities of their neighbors, and instead only depend on the fraction of treated neighbors: \( Y_i(w_i; w_{-i}) = f_i(w_i; \sum_{j \neq i} E_{ij} w_j / \sum_{j \neq i} E_{ij})^2 \), where \( f_i \in \mathcal{F} \) is the potential outcome function of the \( i \)-th subject, which may depend arbitrarily on \( U_i \). We assume the pairs \((U_i, f_i)\) are independent and identically sampled from some distribution on \([0,1] \times \mathcal{F} \).

Relative to the existing literature, the most distinctive assumption we make here is our use of random graph asymptotics. This type of graphon models are motivated by fundamental results on exchangeable arrays (Aldous, 1981; Lovász and Szegedy, 2006), and have received considerable attention in the literature in recent years (e.g., Gao, Lu and Zhou, 2015; Parise and Ozdaglar, 2019; Zhang, Levina and Zhu, 2017); however, we are not aware of previous uses of this assumption to the problem of treatment effect estimation under network interference. For our purposes, working with a graphon model gives us a firm handle on how various estimators behave in the large-sample limit, and opens the door to powerful analytic tools that we will use to prove central limit theorems. In Section 3, we discuss a number of example graphon models in the context of an application.

The anonymous interference assumption was proposed by Hudgens and Halloran (2008) and is commonly used in the literature; Figure 1 illustrates the anonymous interference assumption on a small graph. The specific form of the anonymous interference assumption—where interference only depends on the ratio of treated neighbors but not on the total number of neighbors—is called the “distributional interactions” assumption by Manski (2013).\(^2\)

\(^2\)We take the convention of \(0/0 = 0\).

\(^3\)It is plausible that similar analyses could also be applied to more general cases, e.g., when the potential outcome function is asymptotically additive in the treatments of its neighbors, i.e.,
Given these assumptions, we can characterize our target estimands (1) and (2) in terms of primitives from the graphon sampling model. The following assumption is designed to let us handle both dense graphs, and graphs that are sparse in the sense of Borgs et al. (2019).

**Assumption 4 (Graphon Sequence).** The graphon sequence $G_n(\cdot, \cdot)$ described in Assumption 2 satisfies $G_n(U_i, U_j) = \min \{1, \rho_n G(U_i, U_j)\}$, where $G(\cdot, \cdot)$ is a symmetric, non-negative function on $[0,1]^2$ and $0 < \rho_n \leq 1$ satisfies one of the following two conditions: $\rho_n = 1$ (dense graph), or $\lim_{n \to \infty} \rho_n = 0$ and $\lim_{n \to \infty} n \rho_n = \infty$ (sparse graph). In the case of dense graphs, we simply write $G_n = G$.

Finally, we make an assumption on the smoothness of the potential outcome functions. Intuitively, this assumption states that the potential outcomes do not change much if the fraction of treated neighbors changes a little bit.

**Assumption 5 (Smoothness).** The potential outcome functions $f(w, x)$ satisfy

$$|f(w, x)|, \quad |f'(w, x)|, \quad |f''(w, x)|, \quad |f'''(w, x)| \leq B$$

uniformly in $f \in F$, $w \in \{0, 1\}$ and $x \in [0, 1]$, where all derivatives of $f$ are taken with respect to the second argument.

Proposition 1 provides a simple way of writing down our target estimands in the random graph model spelled out above. Roughly speaking, the direct effect measures how much $f$ changes with its first argument, while the indirect effect is the derivative of $f$ with respect to its second argument. In other words, the direct effect captures the effect of a unit’s own treatment status, while the indirect effect captures the effect of its proportion of treated neighbors. Here—and throughout this paper unless specified otherwise—all proofs are given in the supplementary material.

**Proposition 1.** Consider a randomized trial under network interference satisfying Assumptions 1, 3 and 5, with treatment assigned independently as $W_i \sim \text{Bernoulli}(\pi)$ for some...
0 < \pi < 1. Let \( N_i = \sum_{j \neq i} E_{ij} \) be the number of neighbors of subject \( i \) in the interference graph. Conditional on the interference graph and the potential outcome functions, the estimands (1) can be expressed as follows, where \( B \) is the smoothness constant in (3):

\[
\tau_{\text{DIR}} = \frac{1}{n} \sum_{i=1}^{n} (f_i(1, \pi) - f_i(0, \pi)) + \mathcal{O}\left(\frac{B}{\min_i N_i}\right),
\]

(4)

\[
\bar{\tau}_{\text{IND}} = \frac{1}{n} \sum_{i} \left( \pi f_i'(1, \pi) + (1 - \pi) f_i'(0, \pi) \right) + \mathcal{O}\left(\frac{B}{\sqrt{\min_i N_i}}\right).
\]

Furthermore, if \( \mathbb{E}\left[1/(\min_i N_i)\right] = o(1) \), then the limits taken in (2) exist, and satisfy

\[
\tau_{\text{DIR}} = \mathbb{E}\left[f_i(1, \pi) - f_i(0, \pi)\right], \quad \bar{\tau}_{\text{IND}} = \mathbb{E}\left[\pi f_i'(1, \pi) + (1 - \pi) f_i'(0, \pi)\right].
\]

1.2. Overview of Main Contributions. The key focus of this paper is estimation of the targets (4) and (5) under random graph asymptotics. First, in Section 2, we consider estimation of the direct effect. It is well known there exist simple estimators of \( \bar{\tau}_{\text{IR}} \) that are unbiased under considerable generality and that do not explicitly reference the graph structure \( \{E_{ij}\} \); however, as discussed further in Section 2, getting a sharp characterization of the error distribution of these estimators has proven difficult so far. We add to this line of work by showing that, under our random graph model, these simple estimators in fact satisfy a central limit theorem with \( \sqrt{n} \)-scale errors, regardless of the density of the interference graph as captured by \( \rho_n \). We also provide a quantitative expression for variance inflation due to interference effects in terms of the graphon \( G \).

Next, while the point estimators for the direct effect studied in Section 2 have a simple functional form, the asymptotic variance in the corresponding central limit theorem appears challenging to estimate. To address this challenge, in Section 3 we develop upper bounds for this asymptotic variance that can be used for conservative inference. Our upper bounds are sharp enough to enable meaningful inference in the context of an application, and are robust to having only generic knowledge about the structure of the interference graph \( E \).

Finally, in Section 4, we consider estimation of the indirect effect. This task appears to be substantially more difficult than estimation of the direct effect, and we are aware of no prior work on estimating the indirect effect without either assuming extreme sparsity (e.g., the interference graph has bounded degree), or assuming that the interference graph can be divided up into cliques and that we can exogenously vary the treatment fraction in each clique. Here, we find that natural unbiased estimators for \( \bar{\tau}_{\text{IND}} \) that build on our discussion in Section 2 have diverging variance and are thus inconsistent, even in reasonably sparse graphs. We then propose a new estimator which we call the PC-balancing estimator, and provide both formal and numerical evidence that its error decays as \( \sqrt{\rho_n} \) for sparse interference graphs in the sense of Assumption 4, provided the graphon \( G \) admits low-rank structure.

1.3. Notation. Throughout this paper, we use \( C, C_1, C_2 \ldots \) for constants not depending on \( n \). Note that \( C \) might mean different things in different settings. We let \( f_i'(w, x) \), \( f_i''(w, x) \), etc., denote derivatives of \( f \) with respect to the second argument \( x \). We write \( N_i = \sum_{j \neq i} E_{ij} \) for the number of neighbors of subject \( i \), and \( M_i = \sum_{j \neq i} E_{ij} W_{ij} \) for the number of treated neighbors. We use \( \Omega() \), \( \mathcal{O}() \), \( \Omega_p() \), \( \omega_p() \), \( \sim, \asymp, \ll \) in the following sense: \( a_n = \Omega(b_n) \) if \( a_n \geq C b_n \) for \( n \) large enough, where \( C \) is a positive constant. \( a_n = \mathcal{O}(b_n) \) if \( |a_n| \leq C b_n \) for \( n \) large enough. \( X_n = \Omega_p(b_n) \), if for any \( \delta > 0 \), there exists \( M, N > 0 \), s.t. \( \mathbb{P}[|X_n| \geq M b_n] \leq \delta \) for any \( n > N \). \( X_n = \omega_p(b_n) \), if for any \( \delta > 0 \), there exists \( M, N > 0 \), s.t. \( \mathbb{P}[|X_n| \leq M b_n] \leq \delta \) for any \( n > N \). \( X_n = o_p(b_n) \), if \( \lim \mathbb{P}[|X_n| \geq \epsilon b_n] \rightarrow 0 \) for any \( \epsilon > 0 \). \( a_n \sim b_n \) if \( \lim a_n/b_n = 1 \). \( a_n \asymp b_n \) if there exist \( C > 0 \), s.t. \( \limsup a_n/b_n \leq C \) and
the effect of treatment

\hat{\tau}_{\text{DIR}} = \frac{1}{n} \sum_i W_i Y_i - \frac{1}{n} \sum_i (1 - W_i) Y_i

where as always 0 < \pi < 1 denotes the randomization probability \( W_i \sim \text{Bernoulli}(\pi) \). A simple calculation then verifies that, under interference, the Horvitz-Thompson estimator is unbiased for the \( \bar{\tau}_{\text{DIR}} \) from (1) conditionally on potential outcomes (i.e., conditionally on both the exposure graph and each unit’s response functions):

\[ \mathbb{E} \left[ \hat{\tau}_{\text{DIR}}^\text{HT} \mid Y(\cdot) \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ W_i Y_i(1, W_{-i}) \mid Y_i(\cdot) \right] - \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ (1 - W_i) Y_i(0, W_{-i}) \mid Y_i(\cdot) \right] \]

\[ = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ W_i \mid Y_i(\cdot) \right] \mathbb{E} \left[ Y_i(1, W_{-i}) \mid Y_i(\cdot) \right] - \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ 1 - W_i \mid Y_i(\cdot) \right] \mathbb{E} \left[ Y_i(0, W_{-i}) \mid Y_i(\cdot) \right] = \bar{\tau}_{\text{DIR}}. \]

Sävje, Aronow and Hudgens (2021) use this fact along with concentration arguments to argue that the Horvitz-Thompson estimator is consistent for the direct effect in sparse graphs, with a rate of convergence that depends on the degree of the graph and approaches the parametric 1/\sqrt{n} rate as we push towards a setting where its degree is bounded. Specifically, in their Proposition 2, they argue that

\[ \hat{\tau}_{\text{DIR}} - \bar{\tau}_{\text{DIR}} = O_p \left( \frac{1}{n^2} \sum_{i,j=1}^n H_{ij} \right), \]

where the \( H \) matrix tallies second-order neighbors, i.e., \( H_{ii} = 1 \) for all \( i = 1, \ldots, n \) and for \( i \neq j \) if there exist a node \( k \neq i, j \) such that \( E_{ik} = E_{jk} = 1 \); and \( H_{ij} = 0 \) else.

Here, we revisit the setting of Sävje, Aronow and Hudgens (2021) under our graphon generative model. Our qualitative findings mirror theirs: Familiar estimators of the average
treatment effect without interference remain good estimators of the direct effect from the perspective of random graph asymptotics. However, our quantitative results are substantially sharper. We show that the Horvitz-Thompson estimator is consistent for the direct effect in both sparse and dense graphs, and find that it has a $1/\sqrt{n}$ rate of convergence regardless of the degree of the exposure graph. Furthermore, we establish a central limit theorem for the estimator, and quantify the excess variance due to interference effects.

2.1. A Central Limit Theorem. As discussed above, our goal is to establish that natural estimators of the average treatment effect in the no-interference setting are asymptotically normal around the direct effect once interference effects appear. To this end, we consider both the Horvitz-Thompson estimator (7), and the associated Hájek (or ratio) estimator

$$\hat{\tau}_{\text{DIR}}^{\text{HT}} = \frac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i} - \frac{\sum_{i=1}^{n} (1 - W_i) Y_i}{\sum_{i=1}^{n} 1 - W_i}. \tag{10}$$

Unlike the Horvitz-Thompson estimator, the Hájek estimator is not exactly unbiased; however, its ratio form makes it invariant to shifting all outcomes by a constant.

Our first result is a characterization of the estimators $\hat{\tau}_{\text{DIR}}^{\text{HT}}$ and $\hat{\tau}_{\text{DIR}}^{\text{HAJ}}$ in large samples under the assumption of anonymous interference. This result does not require our graphon generative model, and instead only relies on smoothness of the potential outcome functions $f(w, \pi)$ as well as concentration of quadratic forms of $W_i - \pi$. In particular, this results holds conditionally on the exposure graph and the potential outcome functions.

**Lemma 2.** Under the conditions of Proposition 1 and conditionally on the graph and the potential outcome functions, the estimators of the direct effect defined in (7) and (10) respectively satisfy

$$\hat{\tau}_{\text{DIR}}^{\text{HT}} - \hat{\tau}_{\text{DIR}} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{f_i(1, \pi)}{\pi} + \frac{f_i(0, \pi)}{1 - \pi} \right) (W_i - \pi) + O_p(\delta), \tag{11}$$

$$\hat{\tau}_{\text{DIR}}^{\text{HAJ}} - \hat{\tau}_{\text{DIR}} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{f_i(1, \pi)}{\pi} + \frac{f_i(0, \pi)}{1 - \pi} - \mathbb{E} \left[ \frac{f_i(1, \pi)}{\pi} + \frac{f_i(0, \pi)}{1 - \pi} \right] \right) (W_i - \pi) + O_p(\delta),$$

where $\hat{\tau}_{\text{DIR}}$ is as defined in (1) and

$$\delta = \frac{B}{\sqrt{n \min_i N_i}} + \frac{B}{n \min_i N_i^{3/2}} \left( \sum_{i,j} \gamma_{i,j} \right), \quad \gamma_{i,j} = \sum_{k \neq i,j} E_{ik} E_{jk}. \tag{12}$$

**Proof.** Here we only provide a sketch of proof for the Horvitz-Thompson estimator to illustrate the main idea. The full proof will be given in the supplementary material. To start, as justified by Assumption 5, we can Taylor expand $f_i(w, M_i/N_i)$ into four terms,

$$f_i(w, M_i/N_i) = f_i(w, \pi) + f_i'(w, \pi) (M_i/N_i - \pi) + \frac{1}{2} f_i''(w, \pi) (M_i/N_i - \pi)^2 + r_i(w, M_i/N_i), \tag{13}$$

where
for any \( w \in \{0, 1\} \), where \( r_i(w; M_i/N_i) = \frac{1}{n} f''_i(w_i, \pi^*_i) (M_i/N_i - \pi)^3 \) for some \( \pi^*_i \) between \( \pi \) and \( M_i/N_i \). A careful application of this expansion to both \( \hat{\tau}_{\mathrm{DIR}} \) and \( \hat{\tau}_{\mathrm{HT}} \) establishes that

\[
\hat{\tau}_{\mathrm{HT}} - \hat{\tau}_{\mathrm{DIR}} = \frac{1}{n} \sum_{i=1}^{n} (W_i - \pi) \left( \frac{f_i(1, \pi)}{\pi} + \frac{f_i(0, \pi)}{1 - \pi} \right)
+ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{M_i}{N_i} - \pi \right) \left( f'_i(1, \pi) - f'_i(0, \pi) \right)
+ \frac{1}{n} \sum_{i=1}^{n} (W_i - \pi) \left( \frac{f'_i(1, \pi)}{\pi} + \frac{f'_i(0, \pi)}{1 - \pi} \right)
+ S_1 + S_2 + O_p(\delta),
\]

where \( S_1 \) and \( S_2 \) are as given in (17), and \( \delta \) is as defined in (12); details of the derivation are given in the supplementary material.

We observe that the first summand in (14) matches the first term in (11), while the second summand can be rearranged as follows (while preemptively relabeling the summation index as \( j \)): With \( d_j = f''_j(1, \pi) - f''_j(0, \pi) \), we have

\[
\sum_{j=1}^{n} \left( \frac{M_j}{N_j} - \pi \right) d_j = \sum_{j=1}^{n} \sum_{i \neq j} E_{ij} (W_i - \pi) \frac{d_j}{\sum_{k \neq j} E_{jk} d_j} = \sum_{i=1}^{n} (W_i - \pi) \sum_{j \neq i} \sum_{k \neq j} E_{jk} d_j.
\]

Thus, the first two summands in (14) complete our target expression.

Now, the third summand can be rewritten into a quadratic form in \( W_i - \pi \),

\[
\frac{1}{n} \sum_{i=1}^{n} (W_i - \pi) \left( \frac{M_i}{N_i} - \pi \right) \zeta_i = \frac{1}{n} \sum_{i \neq j} (W_i - \pi) M_{ij} (W_j - \pi),
\]

where \( \zeta_i = f''_i(1, \pi)/\pi + f''_i(0, \pi)/(1 - \pi) \), \( M_{ij} = \zeta_i E_{ij}/N_j \). Since the vector \( W_i - \pi \) has independent and mean-zero entries, we can use the Hanson-Wright inequality as stated in Rudelson and Vershynin (2013) to verify that the above term is bounded in probability to order \( ||M||_F^2/n \), which in turn is bounded as \( O_p(B/\sqrt{n \min_i N_i}) \). It remains to control

\[
S_1 = \frac{1}{2n} \sum_{i=1}^{n} (W_i - \pi) (M_i/N_i - \pi)^2 \left( \frac{f''_i(1, \pi)}{\pi} + \frac{f''_i(0, \pi)}{1 - \pi} \right),
\]

\[
S_2 = \frac{1}{n} \sum_{i=1}^{n} (W_i - \pi) \left( \frac{r_i(1, M_i/N_i)}{\pi} + \frac{r_i(0, M_i/N_i)}{1 - \pi} \right).
\]

Here, both \( S_1 \) and \( S_2 \) have the form of \( \sum_i (W_i - \pi) \alpha_i(M_i/N_i)/n \), where the function \( \alpha_i \) is measurable with respect to \( \{f_j\}_{j=1}^{n} \). We will use Proposition 3 stated below to bound them. In doing so recall that by properties of the Binomial distribution there are constants \( C_k \) such that \( \mathbb{E}[\{(M_i/N_i - \pi)^2 | G, f(\cdot)\}] \leq C_k/N_i^k \) for all \( k = 1, 2, \ldots \). Thus, by Assumption 5, \( \mathbb{E}[(M_i/N_i - \pi)^2 f''_i(1, \pi)^2 | G, f(\cdot)\}] \leq C_2B^2/N_i^2 \) and \( \mathbb{E}[r_i(w, M_i/N_i)/G, f(\cdot)\}] \leq C_3B^2/N_i^3 \), giving us the needed second moment bounds on \( \alpha_i(M_i/N_i) \).

**Proposition 3.** Under the conditions of Lemma 2, let \( \alpha_i : [0, 1] \to \mathbb{R} \) be measurable with respect to \( \{f_j\}_{j=1}^{n} \), and suppose that \( \mathbb{E}[\alpha_i(M_i/N_i)^2 | G, f(\cdot)\}] \leq CB^2/N_i^2 \) almost surely for some universal constant \( C \). Then, conditionally on \( G \) and \( \{f_i(\cdot)\}_{i=1}^{n} \),

\[
\frac{1}{n} \sum_{i=1}^{n} (W_i - \pi) \alpha_i(M_i/N_i) = O_p \left( \frac{B}{\sqrt{n \min_i N_i}} \right).
\]
A sufficient set of conditions for $\delta$ to be negligible is the following: If the minimum degree of the exposure graph is bounded from below as $\min_{i} \{ N_i \} = \Omega_p(n\rho_n)$ and the number of common neighbors $\gamma_{i,j}$ satisfies $\sum_{i,j} \gamma_{i,j} = O_p(n^3\rho_n^3)$, then the $\delta$ term in Lemma 2 obeys $\delta = O_p(B/\sqrt{n^2\rho_n})$. Under our graphon generative model (Assumptions 2 and 4), then (20) and (21) as used in Theorem 4 below imply the above conditions.

The characterization of Lemma 2 already gives us some intuition about the behavior of the estimators of the direct effect. In the setting without interference, it is well known that the Horvitz-Thompson estimator satisfies

$$\hat{\tau}_{\text{HT}} - \bar{\tau} = \frac{1}{n} \sum_{i} \left( \frac{Y_i(1) - Y_i(0)}{\pi} \right) (W_i - \pi),$$

where $\bar{\tau} = \frac{1}{n} \sum_{i=1}^{n} (Y_i(1) - Y_i(0))$ is the sample average treatment effect, and a similar expression is available for the Hájek estimator. Here we found that, under interference, $\hat{\tau}_{\text{HT}}$ preserves this error term, but also acquires a second one that involves interference effects. Qualitatively, the term $\sum_{j\neq i} E_{ij}(f'_i(1, \pi) - f'_j(0, \pi))/\sum_{k\neq j} E_{jk}$, captures the random variation in the outcomes experienced by the neighbors of the $i$-th unit due to the treatment $W_i$ assigned to the $i$-th unit.

It is now time to leverage our graphon generative model. The following result uses this assumption to characterize the behavior of the terms given in Lemma 2, and to establish a central limit theorem that highlights how interference effects play into the asymptotic variation of estimators of the direct effect.

**Theorem 4.** Consider a randomized trial under network interference satisfying Assumptions 1–5, with treatment assigned independently as $W_i \sim \text{Bernoulli}(\pi)$ for some $0 < \pi < 1$. Suppose that the function $g_1(u) := \int_{0}^{1} \min(1, G(u, t))dt$ is bounded away from 0,

$$g_1(u_1) \geq c_1 \text{ for any } u_1,$$

and that the graphon has a finite second moment, i.e.

$$E \left[ G(U_1, U_2)^k \right] \leq c_k^2, \text{ for } k = 1, 2.$$

Finally, suppose that $\liminf \log \rho_n / \log n > -1$. Then, both the Horvitz-Thompson and Hájek estimators have a limiting Gaussian distribution around the direct effect (1),

$$\sqrt{n} (\hat{\tau}_{\text{HT}}^{\text{DIR}} - \bar{\tau}_{\text{DIR}}) \Rightarrow N \left( 0, \pi(1-\pi)E \left[ (R_i + Q_i)^2 \right] \right),$$

$$\sqrt{n} (\hat{\tau}_{\text{HÁJ}}^{\text{DIR}} - \bar{\tau}_{\text{DIR}}) \Rightarrow N \left( 0, \pi(1-\pi) \left( \text{Var} [R_i + Q_i] + (E [Q_i])^2 \right) \right),$$

where

$$R_i = \frac{f_i(1, \pi)}{\pi} + \frac{f_i(0, \pi)}{1-\pi}, \quad Q_i = E \left[ \frac{G(U_i, U_j)(f'_i(1, \pi) - f'_j(0, \pi))}{g(U_j)} \right].$$

If furthermore $\sqrt{n}\rho_n \to \infty$, then similar results hold for the population-level estimand (2), with $\sigma_0^2 = \text{Var} \left[ f_i(1, \pi) - f_i(0, \pi) \right]$:

$$\sqrt{n} (\hat{\tau}_{\text{HT}}^{\text{DIR}} - \bar{\tau}_{\text{DIR}}) \Rightarrow N \left( 0, \sigma_0^2 + \pi(1-\pi)E \left[ (R_i + Q_i)^2 \right] \right),$$

$$\sqrt{n} (\hat{\tau}_{\text{HÁJ}}^{\text{DIR}} - \bar{\tau}_{\text{DIR}}) \Rightarrow N \left( 0, \sigma_0^2 + \pi(1-\pi) \left( \text{Var} [R_i + Q_i] + (E [Q_i])^2 \right) \right).$$

Our first observation is that, in contrast to the upper bounds of Sävje, Aronow and Hudgens (2021), our additional random graph assumptions, paired with anonymous interference and
smoothen, enable us to establish that the asymptotic accuracy of the estimators does not depend on the sparsity level \( \rho_n \). In our setting, direct effects are accurately estimable even in dense graphs, and \( \rho_n \) at most influences second-order convergence to the Gaussian limit.

To further interpret this result, we note that, in the case without interference (i.e., omitting all contributions of \( Q_i \)), the results (22) and (24) replicate well known results about estimators for the average treatment effects. In general, unless \( R_i \) and \( Q_i \) are strongly negatively correlated, then we would expect and \( \mathbb{E} \left[ (R_i + Q_i)^2 \right] > \mathbb{E} \left[ R_i^2 \right] \) and \( \text{Var} \left[ R_i + Q_i \right] + \mathbb{E} \left[ Q_i^2 \right] > \text{Var} \left[ R_i \right] \), meaning that interference effects inflate the variance of both the Horvitz-Thompson and Hájek estimators. However, it is possible to design special problem instances where interference effects in fact reduce variance. The variance inflation between (22) and (24) arises from targeting \( \tau_{\text{DIR}} \) versus \( \tau_{\text{DIR}} \). This corresponds exactly to the familiar variance inflation term that arises from targeting the average treatment effect as opposed to the sample average treatment effect in the no-interference setting; see Imbens (2004) for a discussion. The condition \( \sqrt{n} \rho_n \to \infty \) for (24) is required to make the error term in (4) small.

Theorem 4 also enables us to compare the asymptotics of the Horvitz-Thompson and Hájek estimators. Here, interestingly, the picture is more nuanced. The asymptotic variance of the Horvitz-Thompson estimator depends on \( \mathbb{E} \left[ (R_i + Q_i)^2 \right] = \text{Var} \left[ R_i + Q_i \right] + \mathbb{E} \left[ R_i + Q_i \right]^2 \), and so the Hájek estimator is asymptotically more accurate than the Horvitz-Thompson estimator if and only if \( \mathbb{E} \left[ Q_i^2 \right] \leq \mathbb{E} \left[ R_i + Q_i \right]^2 \). Thus, neither estimator dominates the other one in general. This presents a marked contrast to the case without interference, where the Hájek estimator always has a better asymptotic variance than the Horvitz-Thompson estimator (unless \( \mathbb{E} \left[ R_i \right] = 0 \), in which case they have the same asymptotic variance).

One question left open above is how to estimate the asymptotic variances that arise in Theorem 4, which depends on unknown functionals of \( G \) and the \( f_i \) that may be difficult to estimate.\(^4\) However, even when point estimation of the asymptotic variance is difficult, we may be able to use subject matter knowledge to derive practically useful upper bounds for this asymptotic variance that can be paired with Theorem 4 to build asymptotically conservative confidence intervals for the direct effect. We further investigate this approach below in the context of an application.

2.2. Numerical Evaluation. To validate our findings from Theorem 4, we consider a simple numerical example. Here, we simulate data as described in Section 1.1, for a graph with \( n = 1000 \) nodes generated via a constant graphon \( G_n(u_1, u_2) = 0.4 \), i.e., where any pair of nodes are connected with probability 0.4. We then generate treatment assignments as \( W_i \sim \text{Bernoulli}(\pi) \) with \( \pi = 0.7 \), and potential outcome functions as \( f_i(w, x) = w x / \pi^2 + \varepsilon_i \) with \( \varepsilon_i \sim N(0, 1) \). Figure 2 shows the distribution of the estimators \( \hat{\tau}_{\text{HT}} \) and \( \hat{\tau}_{\text{HJ}} \) and \( \hat{\tau}_{\text{DIR}} \) across \( N = 3000 \) simulations. We see that the distribution of the estimators closely matches the limiting Gaussian distribution from Theorem 4 (in red). In contrast, a simple analysis that ignores interference effects would result in a limiting distribution (shown in blue) that’s much too narrow. In other words, here, ignoring interference would lead one to underestimate the variance of the estimator.

\(^4\)The fundamental difficulty here is that interference creates intricate dependence patterns that break standard strategies for variance estimation. In particular, standard non-parametric bootstrap or subsampling-based methods do not apply here, because removing the \( i \)-th datapoint from the sample does not erase the spillover effects due to the treatment \( W_i \) received by the \( i \)-th person. Thus, in order to estimate the asymptotic variance, it is likely one would need to develop plug-in estimators for the expressions in (24). The main challenge in doing so is with terms of the form \( \mathbb{E}[Q_i^2] \), because the \( Q_i \) involve derivatives of the potential outcome function \( f_i \) with respect to the fraction of treated neighbors. Estimating the first moment of \( Q_i \) appears to be a task whose statistical difficulty is comparable to estimating the indirect effect \( \tau_{\text{IND}} \), while estimating its second moment poses further challenges. We leave a discussion of point-estimators for these quantities to further work.
3. Conservative Intervals for the Direct Effect. Theorem 4 implies that, under our random graph model, accurate estimation of the direct effect is simple and practical. The estimators \( \hat{\tau}_{\text{HT}}^{\text{DIR}} \) and \( \hat{\tau}_{\text{HAJ}}^{\text{DIR}} \) have an elementary functional form, and do not explicitly depend on the interference graph \( E \) and so can be implemented even if we only have incomplete or potentially inaccurate knowledge of it. Using Theorem 4 to build confidence intervals, however, is more challenging. First, the relevant asymptotic variance depends on the graph \( E \) and will be difficult to estimate if we don’t have accurate information about it. Second, as discussed above, even in an ideal setting where \( E \) is known, estimating the asymptotic variance may pose challenges and consistent point estimators are not currently available.

In this section, we explore an alternative approach to using Theorem 4 in practice, based on conservative bounds for unknown components in the asymptotic variance. We illustrate this strategy using a study by Duflo et al. (2013) on how environmental regulations can help curb industrial pollution in Gujarat, India. Our main finding is that, in this example, we can translate reasonably weak assumptions on the high-level structure of the interference graph \( E \) into meaningful bounds on the asymptotic variance of estimators of the direct effect. We believe that similar bounds-based strategies may also be useful in other applications.

Duflo et al. (2013) start from a status quo where the state had specified limits on how much firms may pollute, and industrial plants needed to hire independent auditors to verify compliance. The authors were concerned, however, about a conflict of interest: Because plants hire their own auditors, the auditors may be incentivized to turn a blind eye to potential non-compliance in order to get hired again. To test this hypothesis, Duflo et al. (2013) considered a sample of \( n = 473 \) audit-eligible plants in Gujarat, and randomly assigned half of these plants (i.e., \( \pi = 0.5 \)) to a treatment designed to make auditors work more independently, while the control group remained with the status quo. The treatment had multiple components, including pre-specifying the auditor (instead of letting the plants hire their own auditors), and using a fixed fee rather than a fee negotiated between the plant and the auditor; see Duflo et al. (2013) for details. The authors found a substantial effect of changing the audit mechanism. In particular, they found that plants in the treatment condition reduced combined

---

Fig 2: Histogram of \( \hat{\tau}_{\text{HT}}^{\text{DIR}} \) and \( \hat{\tau}_{\text{HAJ}}^{\text{DIR}} \) across \( N = 3000 \) replications, for a graph of size \( n = 1,000 \). The overlaid curves denote, in red, the limiting Gaussian distribution derived in Theorem 4 and, in blue, the limiting Gaussian distribution we would get while ignoring interference effects.
water and air pollutant emissions by \( \hat{\tau} = 0.211 \) standard deviations of the pollutant emission distribution for control plants, with an associated standard error estimate of 0.99 and a 95% confidence interval \( \tau \in (0.017, 0.405) \).

The analysis used in Duflo et al. (2013) did not consider interference effects, i.e., it assumed that enrolling a specific plant \( i \) in the treatment condition only affected pollution levels for the \( i \)-th plant, not the others. This is, however, a potentially problematic assumption: For example, one might be concerned that some plants are in close contact with each other, and that having one plant be enrolled in the treatment condition would also make some of its closely associated plants reconsider non-compliant pollution.

If we’re worried about interference, how should we reassess the point estimate \( \hat{\tau} = 0.211 \)?

What about the associated confidence interval? As discussed in Section 2, the work of Sävje, Aronow and Hudgens (2021) already provides a good answer to the first question: In the presence of interference, we should understand \( \hat{\tau} = 0.211 \) as an estimate of the direct effect of treatment, while marginalizing over the ambient treatment assigned to other plants. The answer to the second question is more delicate. As shown in Theorem 4, in this case, the width of confidence intervals built around \( \hat{\tau} \) need to be adjusted to account for interference; however, we do not have access to estimators for the variance parameters in (24)—and in fact, here, we don’t even observe the interference graph \( E \). Thus, in order to assess the sensitivity of the findings in Duflo et al. (2013), the best we can hope for is to pair the structure of our result from Theorem 4 with subject-matter knowledge in order to derive conservative bounds for the variance inflation induced by interference.\(^5\)

To this end recall that, under the assumptions of Theorem 4, the Hájek estimator \( \hat{\tau} \) satisfies a central limit theorem\(^6\)

\[
\sqrt{n} (\hat{\tau} - \bar{\tau}_{DIR}) \Rightarrow N(0, \sigma^2 + \pi(1-\pi) V), \quad V = \text{Var} [R_i] + 2 \text{Cov} [R_i, Q_i] + E \left[ Q_i^2 \right] \\
Q_i = Q(U_i) = E \left[ \frac{G(U_i, U_j)(f_j'(1,\pi) - f_j'(0,\pi))}{g(U_j)} \right] | U_i .
\]

Now, let \( V_0 = \text{Var} [R_i] \) measure the asymptotic variance of the Hájek estimator for the average treatment effect that ignores interference effects and note that, by Cauchy-Schwarz,\(^7\)

\[
V \leq V_0 + 2 \sqrt{V_0 E \left[ Q_i^2 \right] + E \left[ Q_i^2 \right].}
\]

\(^5\)Sävje, Aronow and Hudgens (2021) also consider bounds for the variance of \( \hat{\tau} \), but they are not sharp enough to quantitatively engage with the confidence interval of Duflo et al. (2013). More specifically, Sävje, Aronow and Hudgens (2021) propose a number variance estimators that take the form of the product of an inflation factor \( \alpha \) and the baseline variance estimator \( \hat{S} = 1/n^2 (\sum_i W_i Y_i^2 / \pi_2^2 + \sum_i (1 - W_i) Y_i^2 / (1 - \pi_2^2)) \), i.e., they use \( \text{Var}[\hat{\tau}] \leq \alpha \hat{S} \). With this approach, however, the inflation factor \( \alpha \) can be large: If we follow the definition of the \( H \) matrix as in (9), and define \( h_i = \sum_{j \neq i} H_{ij} \) to be the number of second order neighbors, then choices of the inflation factor include the average of \( h_i \), the maximum of \( h_i \), and the largest eigenvalue of the \( H \) matrix. Then, for example, in a simple disjoint-community model where there are 20 communities and roughly 25 plants in each community, the inflation factor would be roughly 25—meaning that confidence interval would need to be widened by a factor of 5 to accommodate interference. Furthermore \( \hat{S} \) itself is conservative even without interference.

\(^6\)The estimate \( \hat{\tau} = 0.211 \) of Duflo et al. (2013) was derived from a linear regression with fixed effects for sub-regions of Gujarat. The regression also included multiple observations per plant, and then clustered standard errors at the plant level. Here, we conduct a sensitivity analysis as though the point estimate \( \hat{\tau} \) had been derived via a Hájek estimator, which is equivalent to linear regression without fixed effects and without multiple observations per plant. It is likely that a sensitivity analysis that also took into account fixed effects and repeated observations would give a similar qualitative picture, but our formal results are not directly applicable to that setting.

\(^7\)One point left implicit here is that standard variance estimators that ignore interference should be seen as estimators of \( \sigma^2 + \pi(1-\pi) V_0 \) in our model. We discuss this point further in the supplementary material, and provide a formal result for the basic plug-in variance estimator that could be used without interference.
Thus, if we use the original standard error estimate from Duflo et al. (2013) for $V_0$, i.e., we set $(\sigma_0^2 + \pi(1-\pi)V_0)/n = 0.099^2$, then bounding the asymptotic variance term in (3) reduces to bounding $\mathbb{E} \left[ Q_i^2 \right]$.

It now remains to develop useful bounds for $\mathbb{E} \left[ Q_i^2 \right]$, in terms of assumptions on both the graphon $G$ and the potential outcome functions $f_i(w, \pi)$. To gain an understanding of the trade-offs at play here, we consider one example where the variance inflation due to interference is exactly zero, one with non-zero but manageable variance inflation, and one where the variance inflation may get out of control easily.

**Example 1 (Additive Interference).** Suppose interference is additive and that units respond to their neighbors treatment in the same way regardless of their own treatment status, i.e., $f_i(w, \pi) = a_i(w) + b_i(\pi)$. Then $f_i'(0, \pi) = f_i'(1, \pi)$ and $\mathbb{E} \left[ Q_i^2 \right] = 0$, meaning that interference has no effect on the asymptotic variance in (3), i.e., $V = V_0$. Thus interference only affects the precision of the Hájek estimator if $f_i(w, \pi)$ is non-additive in its arguments, regardless of the graphon $G$.

**Example 2 (Disjoint Communities Model).** Now suppose that we can divide the graphon $G$ into $k = 1, \ldots, K$ non-overlapping communities, such that the $G$ is the sum of an overall rank-1 term and $K$ community-specific rank-1 terms. More specifically, we assume that there exist intervals $I_k \subset [0, 1]$ such that $G(u, v) = a_0(u)a_0(v) + \sum_{k'=1}^{K} \{1 \{u, v \in I_k\}\} a_k(u)a_k(v)$ for some functions $a_k : I_k \rightarrow [0, 1]$. Given this setting, we can check that for any $u \in I_k$

$g(u) = a_0(u)\bar{a}_0 + a_k(u)\bar{a}_k, \quad \bar{a}_0 = \int_0^1 a_0(v)\,dv, \quad \bar{a}_k = \int_{I_k} a_k(v)\,dv,$

and so

$$Q(u) = \int_0^1 \frac{a_0(u)a_0(v)}{a_0(v)\bar{a}_0 + \sum_{k'=1}^{K} \{1 \{v \in I_{k'}\}\} a_{k'}(v)\bar{a}_{k'}} \mathbb{E} \left[ f_j'(1, \pi) - f_j'(0, \pi) \right] U_j = v$$

$$\leq \frac{a_0(u)}{a_0} \mathbb{E} \left[ f_j'(1, \pi) - f_j'(0, \pi) \right] + \frac{a_k(u)}{a_k} \mathbb{E} \left[ f_j'(1, \pi) - f_j'(0, \pi) \right] U_j \in I_k.$$

Pursuing this line of reasoning and applying Cauchy–Schwarz, we then find that

$$\frac{1}{2} \mathbb{E} \left[ Q_i^2 \right] \leq \mathbb{E} \left[ a^2_0(U_i) \right] \mathbb{E} \left[ a^2_k(U_i) \right] \mathbb{E} \left[ f_j'(1, \pi) - f_j'(0, \pi) \right]^2$$

(26)

$$+ \sum_{k=1}^{K} \mathbb{P} \left[ U_i \in I_k \right] \frac{\mathbb{E} \left[ a^2_0(U_i) \right] \left[ U_i \in I_k \right]}{\mathbb{E} \left[ a^2_k(U_i) \right] \left[ U_i \in I_k \right]} \mathbb{E} \left[ f_j'(1, \pi) - f_j'(0, \pi) \right] \left[ U_i \in I_j \right]^2.$$

In other words, we’ve found that $\mathbb{E} \left[ Q_i^2 \right]$ can be bounded in terms of moments of $f_j'(1, \pi) - f_j'(0, \pi)$ across the disjoint communities, and in terms of the coefficient of variation of the functions $a_k(u)$ that determine the average degree of different nodes.

**Example 3 (Star Graphon).** We end with star-shape interference graphs, and find that they exhibit strong variance inflation due to interference. Pick some small $\eta > 0$ and some $a \in (0, 1)$, and let $G(u, v) = 1 \{u \leq \eta \text{ or } v \leq \eta\} a$. Then $g(u) = a$ if $u \leq \eta$ and $g(u) = \eta a$ if $u > \eta$, meaning that

$$Q(u) = \int_0^\eta \mathbb{E} \left[ f_j'(1, \pi) - f_j'(0, \pi) \right] U_j = v \, dv$$

$$+ \eta^{-1} \{u \leq \eta\} \int_{\eta}^1 \mathbb{E} \left[ f_j'(1, \pi) - f_j'(0, \pi) \right] U_j = v \, dv.$$
Then, in the limit where the nucleus of the “star” gets small, i.e., $\eta \to 0$, we see that
\[
\lim_{\eta \to 0} \eta \mathbb{E} \left[ Q_i^2 \right] = \mathbb{E} \left[ f_i'(1, \pi) - f_i'(0, \pi) \right]^2,
\]
i.e., the variance inflation term diverges at rate $\eta^{-1}$. The reason this phenomenon occurs is that the treatment assignments for a small number units in the nucleus has a large effect on the outcomes of everyone in the system, and this leads to a considerable amount of variance.

In order to study the sensitivity of the findings of Duflo et al. (2013) to interference, we first need to choose some high-level assumptions on $G$ to work with. Here, we move forward in the setting of Example 2, i.e., under the assumption that interference effects is dominated by links between disjoint and unstructured communities. It thus remains to bound the terms in (26). We consider the following:

1. We assume that both the main effects and interference effects are negative (i.e., independent audits reduce pollution overall), and that indirect effects are weaker than the main effects, i.e., $f_i'(1, \pi) - f_i(0, \pi) \leq f_i'(0, \pi), f_i'(1, \pi) \leq 0$, and in particular $\mathbb{E} \left[ f_i'(1, \pi) - f_i'(0, \pi) \right]^2 \leq \tau_{\text{DIR}}^2$.
2. We assume that all terms in (26) that depend on stochastic fluctuations of $a_k(u)$ and $\mathbb{E} \left[ f_i'(1, \pi) \Big| U_i = u \right]$ can be controlled by considering these terms constant and then inflating the resulting bound by a factor 2.

Pooling all this together, we get that
\[
\mathbb{E} \left[ Q_i^2 \right] \leq 8 \tau_{\text{DIR}}^2.
\]

We do not claim that all the steps leading to (27) are all undisputable, but simply that it’s a potentially reasonable starting point for a sensitivity analysis; other subject-matter knowledge may lead to other alternatives to (27) that can be discussed when interpreting results of an application.

Our final goal is the to use this bound on $\mathbb{E} \left[ Q_i^2 \right]$ to see how much we might need to inflate confidence intervals to account for interference. To do so, we proceed by inverting a level-$\alpha$ hypothesis test. Given (25) and (27), the following chi-squared test will only reject with probability at most $\alpha$ under the null-hypothesis $H_0 : \tau_{\text{DIR}} = \tau_0$:
\[
1 \left\{ \left( \frac{\hat{\tau} - \tau_0}{\tau_0} \right)^2 \geq \frac{\Phi(1 - \alpha/2)^2}{n} \left( \sigma_0^2 + \pi (1 - \pi) \left( V_0 + 2 V_0^2 \sqrt{8V_0^2 \tau_0^2 + 8\tau_0^2} \right) \right) \right\}
\]

In our specific case, recall that $\pi = 1/2$ and $n = 473$, and we assumed that $(\sigma_0^2 + \pi (1 - \pi) V_0)/n = 0.099^2$. This leaves the relationship between $\sigma_0^2$ and $V_0$ unspecified; however, we can maximize the noise term in (28) by setting $\sigma_0^2 = 0$ and $V_0 = 4 \times 473 \times 0.99^2$, which is what we do here, resulting in a fully specified hypothesis test,
\[
1 \left\{ \left( \frac{\hat{\tau} - \tau_0}{\tau_0} \right)^2 \geq \Phi(1 - \alpha/2)^2 \left( 0.0098 + 0.0129 \left| \tau_0 \right| + 0.0042 \tau_0^2 \right) \right\},
\]

which we can now invert.

By applying this strategy, we obtain the following interference-robust 95% confidence interval: $\tau_{\text{DIR}} \in (0.015, 0.464)$. Recall that, in contrast, the unadjusted Gaussian confidence interval was $(0.017, 0.405)$. Interestingly, while our interference adjustment noticeably increased the upper endpoint of this interval, it barely touched the lower endpoint at all; and, as a consequence of this, we are still able to reject the null that $\tau_{\text{DIR}} = 0$ at the 95% level. Figure 3 shows the intervals obtained by inverting (29) for different significance levels $\alpha$.

The reason our confidence intervals are less sensitive to interference as we approach 0 is that we assumed above that indirect effects should be bounded on the order of direct effects;
Fig 3: Level-\(\alpha\) confidence intervals for the direct effect \(\tau_{\text{DIR}}\) in the setting of Duflo et al. (2013). The dashed blue lines denote upper and lower endpoints of a basic Gaussian confidence interval with standard error estimate 0.099, while the solid red curves denote intervals endpoints of a confidence interval derived by inverting (29). The solid line at \(\tau = 0.211\) denotes the point estimate.

thus, when testing a null hypothesis that direct effects are very small, the variance inflation due to indirect effects should also be small. If one were to make different assumptions (e.g., that indirect effects may be large even when direct effects are small), a sensitivity analysis might lead to different conclusions.

A thorough sensitivity analysis of the robustness of the results of Duflo et al. (2013) to interference would also involve examining different assumptions on the graphon, etc., and comparing findings across settings. However, we hope that our discussing so far has helped highlight the promise of using random graph analysis to quantitatively and usefully assess robustness of treatment effect estimators to potential variance inflation due to interference.

4. Estimating the Indirect Effect. We now consider estimation of the indirect effect, i.e., how a typical unit responds to a change in its fraction of treated neighbors. There is some existing literature on this task; however, it has mostly focused on a setting where one has access to many independent networks (Baird et al., 2018; Basse and Feller, 2018; Hudgens and Halloran, 2008; Tchetgen Tchetgen and VanderWeele, 2012). This is also referred to as a partial interference assumption, which states that there are disjoint groups of units and spillover across groups is not allowed. Then, the total effect—and thus also the indirect effect—can be identified by randomly varying the treatment probability \(\pi\) across different groups and regressing the mean outcome in each group against its treatment probability.

Meanwhile, in the single network setting, we note a recent paper by Leung (2020), who studies estimation of both direct and indirect effects when the degree of the exposure graph remains bounded as the sample size \(n\) gets large. He then proposes an estimator that is consistent and has a \(1/\sqrt{n}\) rate of convergence. At a high level, the motivating insight behind his approach is that, in the case of a bounded-degree interference graph, we’ll be able to see
infinitely many (linear in \( n \)) units for any specific treatment signature consisting of the number of neighbors, the number of treated neighbors and the treatment allocation. Hence we can take averages of outcomes with a given treatment signature, and use them to estimate various causal quantities. This strategy, however, does not seem to be extensible to denser graphs.

Our goal here is to develop methods for estimating the indirect effect that can work with a single network that is much denser than those considered by Leung (2020), i.e., \( n \rho_n \gg 1 \) following Assumption 4. We are not aware of any existing results in this setting. Our main contribution is an estimator, the PC balancing estimator, that can be used to estimate the indirect effect in a setting where the graphon \( G \) admits a low-rank representation, i.e., \( G(u, v) = \sum_{k=1}^{r} \lambda_k \psi_k(u) \psi_k(v) \) for a small number \( r \) of measurable functions \( \psi_k : [0, 1] \to \mathbb{R} \). We prove that our estimator converges to the indirect effect at rate \( \sqrt{\rho_n} \) and satisfies a central limit theorem. At a high level, the reason we are able to consistently estimate the indirect effect from a single graph is that, even with reasonably dense graphs, some units will have a higher proportion of treated neighbors than others due to random fluctuations in the treatment assignment mechanism—and our graphon generative assumptions enable us to carefully exploit this variation for consistent estimation.

4.1. An Unbiased Estimator. We start by discussing a natural unbiased estimator for the indirect effect that starts from a simple generalization of Horvitz-Thompson weighting. Recall that the total effect is

\[
\bar{\tau}_{\text{TOT}}(\pi) = \frac{d}{d\pi} \bar{V}(\pi), \quad \bar{V}(\pi) = \frac{1}{n} \sum_i E_{\pi}[Y_i|Y(\cdot)].
\]

For any \( \pi' \in (0, 1) \), the Horvitz-Thompson estimate of \( \bar{V}(\pi') \) is

\[
\hat{V}(\pi') = \frac{1}{n} \sum_{i=1}^{n} Y_i \left( \frac{\pi'}{\pi} \right)^{M_i + W_i} \left( \frac{1 - \pi'}{1 - \pi} \right)^{(N_i - M_i) + (1 - W_i)},
\]

where as usual \( M_i \) is the number of treated units and \( N_i \) the number of neighbors.\(^8\) Thus, as \( \bar{\tau}_{\text{TOT}}(\pi) \) is the derivative of \( \bar{V}(\pi) \), one natural idea is to estimate \( \bar{\tau}_{\text{TOT}}(\pi) \) by taking the derivative of \( \hat{V}(\pi') \):

\[
\bar{\tau}_{\text{TOT}}^U(\pi) = \left[ \frac{d}{d\pi'} \hat{V}(\pi') \right]_{\pi' = \pi} = \frac{1}{n} \sum_i Y_i \left( \frac{M_i + W_i}{\pi} - \frac{N_i - M_i + 1 - W_i}{1 - \pi} \right).
\]

One can immediately verify that this estimator is unbiased for \( \bar{\tau}_{\text{TOT}}(\pi) \) (hence the superscript \( U \)) by noting that \( \hat{V}(\pi') \) is unbiased for \( \bar{V}(\pi') \) following the line of argumentation used in (8). We also note that unbiasedness in (32) follows immediately from the argument of Stein (1981) applied to the binomial distribution. Next, the unbiased estimator of the total effect can be naturally decomposed into two parts:

\[
\bar{\tau}_{\text{TOT}}^U = \frac{1}{n} \sum_i Y_i \left( \frac{M_i + W_i}{\pi} - \frac{N_i - M_i + 1 - W_i}{1 - \pi} \right) = \bar{\tau}_{\text{DIR}}^U + \bar{\tau}_{\text{IND}}^U
\]

\(^8\)One might also be tempted to study the problem of off-policy evaluation in our setting, i.e., using notation from (30), estimating \( V(\pi') \) for \( \pi' \neq \pi \). This, however, appears to be a difficult problem outside of very sparse graphs. For example, in Proposition 9 of the first arXiv version of this paper, we showed the estimator (31) for \( \hat{V}(\pi') \) diverges in a random graph model whenever its average degree grows faster than \( \log(n) \).
Recalling that $\tilde{z}^{HT}_{\text{DIR}}$ is unbiased for $\tau_{\text{DIR}}$, we see that $\tilde{z}^{U}_{\text{IND}}$ is also unbiased for $\tau_{\text{IND}}$.

Unfortunately, however, despite its simple intuitive derivation and its unbiasedness, this estimator is not particularly accurate. More specifically, as shown below, its variance goes to infinity as $n \to \infty$ wherever $\sqrt{n}\rho_n \to \infty$; in other words, this estimator is inconsistent even if most units in the graph only share edges with a fraction $1/\sqrt{n}$ of other units.

**Proposition 5.** Let $\nu = \mathbb{E} \left[ (\pi f_1(1, \pi) + (1 - \pi) f_1(0, \pi)) g(U_1) \right]^2$. If $\sqrt{n}\rho_n \to \infty$, then under the conditions of Theorem 4, $\text{Var} \left[ \tilde{z}^{U}_{\text{IND}} \right] \sim \nu n\rho_n^2$ and $\text{Var} \left[ \tilde{z}^{U}_{\text{TOT}} \right] \sim \nu n\rho_n^2$.

4.2. The PC-Balancing Estimator. In order to develop a new estimator effect for the indirect effect that is robust to the variance explosion phenomenon documented in Proposition 5, we focus on a setting where the graphon is low rank with rank $r$, i.e., our graphon can be written in a form of

$$G(U_i, U_j) = \sum_{k=1}^{r} \lambda_k \psi_k(U_i) \psi_k(U_j)$$

for some function $\psi_k$. The low-rank condition (34) quantifies an assumption that each unit can be characterized using a small number ($r$) of factors, and that the probability of edge formation between two units is a bilinear function of both of their factors. For example, in a social network, we may assume that the probability of two people becoming friends is explained by a few factors including their education, experience, and personality. Such low-rank factor models are a popular way of capturing unobserved heterogeneity; see Athreya et al. (2017) for a recent discussion and references.

Now, in order to develop a consistent estimator, we first need to understand why the unbiased estimator fails. To this end, consider a simple stochastic block model with $r$ communities, where any two units in the same community are connected with probability $\rho_n$ while units in different communities are never connected. Letting $k(i)$ denote the $i$-th unit’s community, we can re-write (33) as

$$\tilde{z}^{U}_{\text{IND}} = \frac{1}{\pi(1 - \pi)n} \sum_i \left( \mu_k(i) + (Y_i - \mu_k(i)) \right) \left( M_i - \pi N_i \right)$$

(35) $$= \frac{1}{\pi(1 - \pi)n} \left( \sum_{k=1}^{r} \mu_k \left( \sum_{i : k(i) = k} \left( M_i - \pi N_i \right) \right) + \sum_i \left( Y_i - \mu_k(i) \right) \left( M_i - \pi N_i \right) \right),$$

where $\mu_k = \mathbb{E}_\pi \left[ Y_i \mid k(i) = k \right]$ is the expected outcome in the $k$-th community under our sampling model. In the above expression, the first term is problematic. Specifically

$$\sum_{k(i) = k} \left( M_i - \pi N_i \right) = \sum_{k(i), k(j) = k} E_{ij} (W_j - \pi) = \sum_{k(j) = k} N_j (W_j - \pi)$$

is mean zero, but has variance of scale $\alpha_k n\rho_n^2$, where $\alpha_k$ denotes the fraction of units in community $k$. In other words, in (35), the first term is a major source of noise but contains no information about the indirect effect. In contrast, all the useful information is contained in the second term. It has a non-zero (non-vanishing) mean and is of constant scale. Any successful adaptation of this estimator must thus find a way to effectively cancel out this first term while preserving the second.

Now, given this observation, we can readily mitigate the problematic first term in the context of the stochastic block model considered in (35); for example, we could get rid of it centering the outcomes $Y_i$ in each community $k = 1, \ldots, r$ before running (33). The main question is in how to adapt this insight and remedy to more general specifications.
beyond the stochastic block model. To this end, recall that the stochastic block model considered above is a special case of our setting as spelled out in Assumption 4, where the interval $[0, 1]$ has been partitioned into $r$ non-overlapping sets $I_k$, the $i$-th unit is in community $k$ whenever $u \in I_k$, and the graphon $G$ has a rank-$r$ representation (34) with $\psi_k(u) = \{u \in I_k\}/\sqrt{P[U_i \in I_k]}$ and $\lambda_k = P[U_i \in I_k]$. We also note that the problematic noise term in (35) shows up whenever $E[Y_i \psi_k(U_i)] \neq 0$.

Given this observation, it’s natural to conjecture that if $G$ is any graphon that admits a low-rank representation as in (34), then modifying the unbiased estimator (33) in a way that projects out signal components that are correlated with the eigencomponents $\psi_k(u)$ of the graphon will result in a consistent estimator.

Our proposed PC balancing estimator is motivated by this insight. For simplicity, we start by presenting an “oracle” version of our estimator that assumes a-priori knowledge of the eigencomponents $\psi_k(u)$ of the graphon. The unbiased estimator (33) belongs to a class of weighted estimators $\sum_i \gamma_i Y_i$. We would be able to avoid any noise from signal components associated with the $\psi_k(u)$ if we could modify the weights $\gamma_i$ such that they balance out the $\psi_k(U_i)$ functions, i.e., if $\sum_i \gamma_i \psi_k(U_i) = 0$ for all $k = 1, \ldots, r$. The oracle PC balancing estimator achieves this goal by simply projecting out the relevant parts of the weights as follows:

$$\hat{\psi}_{PC_{IND}} = \frac{1}{n} \sum_i Y_i \left( \frac{M_i}{\pi} - \frac{N_i - M_i}{1 - \pi} + \sum_{k=1}^r \hat{\beta}_k \psi_k(U_i) \right), \text{ where } \hat{\beta} \text{ solves}$$

$$\sum_i \psi_k(U_i) \left( \frac{M_i}{\pi} - \frac{N_i - M_i}{1 - \pi} + \sum_{k=1}^r \hat{\beta}_k \psi_k(U_i) \right) = 0, \text{ for all } l = 1, 2, \ldots, r.$$  

(36)

Now, in practice of course the graphon is usually unknown and we don’t have access to $\psi_k(U_i)$ directly. But if the graphon is low rank as in (34), then the edge probability matrix $[G_n(U_i, U_j)]$ will also be a low rank matrix (when $G_n = \rho_n G$), and its eigenvectors are approximately $\psi_k(U_i)$. We also note that the adjacency matrix $E$ is a noisy observation of this low rank edge probability matrix. Hence, we can estimate $\psi_k(U_i)$ using the eigenvectors $\hat{\psi}_{ki}$ of $E$, and then use the data-driven $\hat{\psi}_{ki}$ to obtain a feasible analogue to (36). We summarize the resulting PC balancing algorithm as Procedure 1. Note that unlike the estimators considered in Section 2, the PC balancing algorithm requires knowledge of the graph $E$.

Our main formal result about the indirect effect establishes consistency and asymptotic normality of the PC balancing estimation in the “sparse” graph setting (i.e., with $\rho_n \to 0$).

We state our result in terms of Bernstein’s condition: Given a random variable $X$ with mean $\mu = E[X]$ and variance $\sigma^2 = E[X^2] - \mu^2$, we say that Bernstein’s condition with parameter $b$ holds if

$$E[(X - \mu)^k] \leq \frac{1}{2} k! \sigma^2 b^{k-2} \quad \text{for } k = 2, 3, 4, \ldots.$$  

Wainwright (2019) shows that one sufficient condition for Bernstein’s condition to hold is that $X$ be bounded. The proof of the following result is given in Section 4.3 below.

**Theorem 6.** Under the conditions of Theorem 4, assume furthermore that we have a sparse graph such that

$$\lim \inf \frac{\log \rho_n}{\log n} > -\frac{1}{2} \quad \text{and} \quad \lim \sup \frac{\log \rho_n}{\log n} < 0.$$  

---

9Throughout this paper, we assume that the rank $r$ of the graphon is known. In practice, one could estimate $r$ by thresholding the eigenvalues of the adjacency matrix using, e.g., the approach of Chatterjee (2015).
Finally suppose that we have a rank-$r$ graphon of the form (34) such that
\[ |\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_r| > 0, \quad \mathbb{E} \left[ \psi_k(U_i)^2 \right] = 1, \quad \text{and } \mathbb{E} \left[ \psi_k(U_i) \psi_l(U_j) \right] = 0 \text{ for } k \neq l, \]
and that for $U_1, U_2 \overset{i.i.d.}{\sim}$ Uniform[0, 1],
\[ \psi_k(U_i) \text{ satisfies the Bernstein condition (39) with parameter } b. \]
Then the PC balancing estimator satisfies
\[ \frac{\hat{\tau}_{\text{PC}}^\text{IND} - \tau_{\text{IND}}}{\sqrt{p_n \sigma_{\text{IND}}^2}} \Rightarrow \mathcal{N}(0, 1), \quad \frac{\hat{\tau}_{\text{PC}}^\text{IND} - \hat{\tau}_{\text{IND}}}{\sqrt{p_n \sigma_{\text{IND}}^2}} \Rightarrow \mathcal{N}(0, 1), \]
where $\sigma_{\text{IND}}^2 = \mathbb{E} \left[ G(U_1, U_2) (\alpha_i^2 + \alpha_1 \alpha_2) \right] + \mathbb{E} \left[ g(U_1) \eta_i^2 \right] / (\pi (1 - \pi)), \quad \alpha_i = f_i(1, \pi) - f_i(0, \pi), \quad b_i = \pi f_i(1, \pi) + (1 - \pi) f_i(0, \pi) \quad \text{and } \eta_i = b_i - \sum_{k=1}^r \mathbb{E} \left[ b_i \psi_k(U_i) \right] \psi_k(U_i).$

As discussed above, we are not aware of any previous results that allow for consistent estimation of the indirect effect in generic, moderately sparse graphs. Here, to establish (43), we need the graph to be “sparse” in the sense that the average fraction of units that are connected decays as $p_n \to 0$; however, we still allow the average degree of the graph to grow very large. In contrast, existing results (e.g., Leung, 2020) require the degree distribution to remain constant, which would amount to setting $p_n \sim 1/n$ in our setting. We also note that the rate of convergence derived for our estimator, namely $\sqrt{p_n}$, is worse than the $1/\sqrt{n}$ we obtained for the direct effect in Theorem 4; however, this appears to be a consequence of the intrinsic difficulty of the task of estimating the indirect effect as opposed to the direct effect.

Finally, given our result for indirect effect, we can naturally get similar results for the total effects. Define $\hat{\tau}_{\text{TOT}}^\text{PC} = \hat{\tau}_{\text{IND}}^\text{PC} + \hat{\tau}_{\text{DIR}}^\text{HT}$. Note that by Theorem 4, $\text{Var} \left[ \hat{\tau}_{\text{DIR}}^\text{HT} \right] = \mathcal{O}_p(1/n) \ll p_n$. Hence a central limit theorem for $\hat{\tau}_{\text{TOT}}^\text{PC}$ can be obtained as well.
Corollary 7. Under the conditions of Theorem 6, the PC balancing estimator is asymptotically normally distributed around the total effect

\[ \frac{\hat{\tau}_{PC} - \tau_{TOT}}{\sqrt{\rho_n \sigma_{IND}}} \Rightarrow \mathcal{N}(0, 1), \quad \frac{\hat{\tau}_{PC}^2 - \tau_{TOT}^2}{\sqrt{\rho_n \sigma_{IND}}} \Rightarrow \mathcal{N}(0, 1), \]

where \( \sigma_{IND} \) is defined the same way as in Theorem 6.

4.3. Proof of Theorem 6. As a preliminary to proving our central limit theorem for \( \hat{\tau}_{PC} \), we need to characterize the behavior of the eigenvectors \( \hat{\psi}_{ki} \) as estimators of the graphon eigenfunctions \( \psi_k(U_i) \), and to show that if we choose \( \hat{\beta} \) to cancel out noise in the direction of \( \hat{\psi}_{ki} \) using (38), then we also effectively balance out signal in the direction of \( \psi_k(U_i) \). Our main tool for doing so is the following lemma. In order to facilitate the interpretation of \( \hat{\psi}_{ki} \) as an estimate of \( \psi_k(U_i) \), in the result below (and throughout this proof), we normalize eigenvectors so that \( \|\psi_k\|_2, \|\psi_k\|_2^2 = n \). Let \( \hat{\Psi} = [\hat{\psi}_1, \ldots, \hat{\psi}_r] \) be the \( n \times r \) matrix whose \( k \)-th column is \( \hat{\psi}_k \).

Lemma 8. Under Assumptions 1, 2 and 4, suppose furthermore that (40), (41) and (42) hold. Let \( \psi_k \) denote the vector of \( \psi_k(U_i) \). There exists an \( r \times r \) orthogonal matrix \( \hat{R} \), such that if we write \( \hat{\Psi}^R = \hat{\Psi} \hat{R} \), and let \( \hat{\psi}_k^R \) be the \( k \)-th column of \( \hat{\Psi}^R \), then, for any vector \( a \) that is independent of \( E \) given \( U_i \)'s, we have

\[ \left| a^T (\hat{\psi}_k^R - \psi_k) \right| / \|a\|_2 = O_p(1). \]

Qualitatively, the above result guarantees that stochastic fluctuations in \( \hat{\psi}_k \) aren’t systematically aligned with any specific vector \( a \); and so, when studying \( \hat{\tau}_{PC} \), the fact that we target \( \hat{\psi}_j \) in (38) shouldn’t induce too much bias. Formally, it is related to the classical result of Davis and Kahan (1970) on the behavior of eigenvectors of a random matrix (and, in our proof, we rely on a variant of the Davis-Kahan theorem given in Yu, Wang and Samworth (2015)). We also note that, given our normalization of \( \|\psi_k\|_2 \), the error bound in (45) is fairly strong—and this type of result is needed in our proof. For example, recent work by Abbe et al. (2020) provides sup-norm bounds on the fluctuations of \( \hat{\psi}_k \); however, these bounds do not decay fast enough to be helpful here.

We are now ready to study \( \hat{\tau}_{PC} \) itself. To this end, we start by decomposing the estimator into parts using the Taylor expansion as justified by (3):

\[
\hat{\tau}_{PC} = \frac{1}{n} \sum_i \left( \frac{M_i}{\pi} - \frac{N_i - M_i}{1 - \pi} + \sum_{k=1}^r \hat{\beta}_k \hat{\psi}_{ki} \right) Y_i
\]
\[
= \frac{1}{n} \sum_i \left( \frac{M_i}{\pi} - \frac{N_i - M_i}{1 - \pi} + \sum_{k=1}^r \hat{\beta}_k \hat{\psi}_{ki} \right) f_i(W_i, \pi)
\]
\[
+ \frac{1}{n} \sum_i \left( \frac{M_i}{\pi} - \frac{N_i - M_i}{1 - \pi} \right) \left( \frac{M_i}{N_i} - \pi \right) f_i'(W_i, \pi)
\]
\[
+ \frac{1}{n} \sum_i \left( \sum_{k=1}^r \hat{\beta}_k \hat{\psi}_{ki} \right) \left( \frac{M_i}{N_i} - \pi \right) f_i''(W_i, \pi)
\]
\[
+ \frac{1}{n} \sum_i \left( \frac{M_i}{\pi} - \frac{N_i - M_i}{1 - \pi} + \sum_{k=1}^r \hat{\beta}_k \hat{\psi}_{ki} \right) \left( \frac{M_i}{N_i} - \pi \right)^2 f_i'''(W_i, \pi^*_i),
\]

\[ n \to \infty, \quad n \log n \to \infty. \]
where \( \pi^*_i \) is some value between \( \pi \) and \( M_i/N_i \). This decomposition already provides some insight into the behavior of \( \hat{\tau}^{PC}_\mathrm{IND} \). Here, the second summand is the one that contains all the signal, while the third and fourth end up being negligible. In particular, we note that the error terms in all three bounds below are smaller than the leading-order \( \sqrt{p_n} \) error in (43).

**Proposition 9.** Under the conditions of Theorem 6,

\[
\frac{1}{n} \sum_i \left( \frac{M_i}{\pi} - \frac{N_i - M_i}{1 - \pi} \right) \left( \frac{M_i}{N_i} - \pi \right) f'_i(W_i, \pi) = \tau_{\mathrm{IND}} + O_p(Bp_n).
\]

**Proposition 10.** Under the conditions of Theorem 6,

\[
\frac{1}{n} \sum_i \left( \frac{r \beta_k \hat{\psi} \hat{\psi}_{ki}}{\pi} \right) \left( \frac{M_i}{N_i} - \pi \right) f'_i(W_i, \pi) = O_p(Bp_n).
\]

**Proposition 11.** Under the conditions of Theorem 6,

\[
\frac{1}{n} \sum_i \left( \frac{M_i}{\pi} - \frac{N_i - M_i}{1 - \pi} + \sum_{k=1}^r \beta_k \hat{\psi}_{ki} \right) \left( \frac{M_i}{N_i} - \pi \right)^2 f''_i(W_i, \pi^*_i) = O_p \left( \frac{B}{\sqrt{p_n}} \right).
\]

It now remains to study the first term in (46). It is perhaps surprising at first glance that this term matters much, since it has nothing to do with cross-unit interference. However, this term ends up being the dominant source of noise; and, in fact, is also what causes the variance of the unbiased estimator \( \hat{\tau}^{U}_{\mathrm{IND}} \) to explode as seen in Proposition 5.

To this end, we introduce some helpful notation. Let \( b_i = \pi f_i(1, \pi) + (1 - \pi) f_i(0, \pi) \), and let \( \mu_k \) be the projection of \( b_i \) onto \( \psi_k(U_i) \), i.e., \( \mu_k = \mathbb{E} [b_i \psi_k(U_i)] \) (recall that \( \mathbb{E} [\psi_k^2(U_i)] = 1 \)). Then, we can express \( f_i(W_i, \pi) \) as

\[
f_i(W_i, \pi) = (W_i - \pi) [f_i(1, \pi) - f_i(0, \pi)] + (\pi f_i(1, \pi) + (1 - \pi) f_i(0, \pi)) \]

\[
= (W_i - \pi) [f_i(1, \pi) - f_i(0, \pi)] + \sum_{k=1}^r \mu_k \psi_k(U_i) + \eta_i,
\]

where \( \eta_i \) is the residual term implied by the above notation. The key property of this decomposition is that, because \( \mu_k \) capture the projection of \( b_i \) onto the \( \psi_k(U_i) \), then \( \mathbb{E} [\eta_i \psi_k(U_i)] = 0 \) for all \( k = 1, \ldots, r \).

Following the discussion around (35) if we did not use the PC balancing adjustment, the problematic term in (50) would be the second one, i.e., the one that’s aligned with the \( \psi_k(U_i) \). But the PC balancing adjustment helps mitigate the behavior of this term. Specifically, thanks to (38), we see that in the context of the first summand of (46),

\[
\frac{1}{n} \sum_i \left( \frac{M_i}{\pi} - \frac{N_i - M_i}{1 - \pi} + \sum_{k=1}^r \beta_k \hat{\psi}_{ki} \right) \sum_{k=1}^r \mu_k \psi_k(U_i)
\]

\[
= \frac{1}{n} \sum_{k=1}^r \mu_k \sum_i \left( \frac{M_i}{\pi} - \frac{N_i - M_i}{1 - \pi} + \sum_{k=1}^r \beta_k \hat{\psi}_{ki} \right) \left( \psi_k(U_i) - \hat{\psi}^R_{ki} \right),
\]

i.e., this term gets canceled out to the extent that \( \hat{\psi}^R_{ki} \) acts as a good estimate of \( \psi_k(U_i) \). The following result, which makes heavy use of Lemma 8 given above, validates this intuition.

**Proposition 12.** Under the conditions of Theorem 6, (51) is bounded as \( O_p(Bp_n) \).
We are now essentially ready to conclude. By combining Propositions 9–12 above and plugging (51) into (46), we can verify the following using basic concentration arguments. In doing so, we heavily rely on the fact that \( \mathbb{E}[\eta_i \psi_k(U_i)] = 0 \), which implies that terms of the type \( \sum_i \eta_i \psi_k(U_i) \) are small.

**Proposition 13.** Under the conditions of Theorem 6,

\[
\hat{\tau}_{\text{IND}}^{\text{PC}} - \tau_{\text{IND}} = \frac{1}{n\pi(1 - \pi)} \sum_{(i,j), i \neq j} (W_i - \pi)E_{ij} \xi_j + o_p(\sqrt{n})
\]

(52)

\[
\xi_j = (W_j - \pi) (f_i(1, \pi) - f_i(0, \pi)) + \eta_j.
\]

It now remains to prove a central limit theorem for the asymmetric bilinear statistic appearing in the right-hand side of (52). To do so, we rely on a central limit theorem for the average of locally dependent random variables derived in Ross (2011) via Stein’s method for Gaussian approximation. The following result leads to our desired conclusion regarding convergence around \( \tau_{\text{IND}} \).

**Proposition 14.** Under the conditions of Theorem 6 and using notation from (51), \( \epsilon_n = (n\pi(1 - \pi))^{-1} \sum_{i \neq j} (W_i - \pi)E_{ij} \xi_j \) has a Gaussian limiting distribution:

\[
\frac{\epsilon_n}{\sqrt{\rho_n}} \Rightarrow \mathcal{N}(0, \sigma_{\text{IND}}^2), \quad \alpha_i = f_i(1, \pi) - f_i(0, \pi),
\]

(53)

\[
\sigma_{\text{IND}}^2 = \mathbb{E}[G(U_1, U_2) \left( \alpha_1^2 + \alpha_1 \alpha_2 \right)] + \mathbb{E}[g(U_1)\eta_1^2] / (\pi(1 - \pi)).
\]

Finally, regarding \( \tilde{\tau}_{\text{IND}} \), we note that \( \tau_{\text{IND}} - \tilde{\tau}_{\text{IND}} \) is an average of i.i.d. random variables bounded by \( CB \). Hence \( \tilde{\tau}_{\text{IND}} = \tau_{\text{IND}} + O_p(B/\sqrt{n}) \), and so the same central limit theorem holds if we center our estimator and \( \tilde{\tau}_{\text{IND}} \) instead.

4.4. Numerical Evaluation. We end this section by empirically evaluating the above findings. First, we evaluate the scaling of the mean-squared error (MSE) of different estimators of the indirect effect. In Figures 4a and 4b, we plot the log-MSE of our PC balancing estimator \( \hat{\tau}_{\text{IND}}^{\text{PC}} \) against the log sample size \( \log(n) \) in a variety of settings described in the supplementary material. In Figure 4a, we consider specifications with sparsity level \( \rho_n = n^{-1/5} \), while in Figure 4b, we consider \( \rho_n = n^{-2/5} \). Theorem 6 predicts that the MSE of \( \hat{\tau}_{\text{IND}}^{\text{PC}} \) should scale as \( \rho_n \), and here, in line with this prediction, we see that the curves in Figures 4a and 4b are roughly linear with slopes \(-1/5 \) and \(-2/5 \) respectively. Next, in Figures 5a and 5b, we perform the same exercise with the unbiased estimator \( \hat{\tau}_{\text{IND}}^U \). By Proposition 5, we know that the MSE of this estimator scales as \( \eta \rho^2 \), and so we expect to see linear relationships with a slope of \( 3/5 \) when \( \rho_n = n^{-1/5} \) and \( 1/5 \) when \( \rho_n = n^{-2/5} \). The slope of the realized MSE is again aligned with the prediction from theory. Finally, we evaluate the predicted distribution for our PC balancing estimator \( \hat{\tau}_{\text{IND}}^{\text{PC}} \) on a larger simulation setting: We consider a rank-3 stochastic block model and a graph with \( n = 1,000,000 \) nodes. Figure 6 shows the distribution of \( \hat{\tau}_{\text{IND}}^{\text{PC}} \) across \( N = 1000 \) simulations. We see that the distribution of the estimator closely matches the limiting Gaussian distribution predicted by Theorem 6.

5. Discussion. The network interference model is a popular framework for studying treatment effect estimation under cross-unit interference. In this paper, we studied estimation in the network interference model under random graph assumptions and showed that—when paired with conditions such as anonymous interference—these assumptions could be leveraged to provide strong performance guarantees. We considered estimation of both the direct
and indirect effects and, for the former, found that existing estimators can be much more accurate than previously known while, for the latter, we proposed a new estimator that is consistent in moderately dense settings. Both sets of results highlight the promise of random graph asymptotics in yielding insights about the nature of treatment effect estimation under network interference and in providing guidance for new methodological developments.

The finding from Theorem 4 that natural estimators of the direct effect satisfy a $1/\sqrt{n}$-rate central limit theorem even in dense graphs may prove to be of particular practical interest. This is because, as emphasized in Sävje, Aronow and Hudgens (2021), the considered estimators of the direct effect are algorithmically the same as standard estimators of the average treatment effect in a randomized study without interference, and so our result for the direct effect can be used to assess the sensitivity of randomized study inference to the presence of unknown network interference. Our $1/\sqrt{n}$-rate guarantees are much stronger than the generic bounds given in Sävje, Aronow and Hudgens (2021), and thus paint a more optimistic picture of how badly unknown interference may corrupt randomized study inference.

One question left open by this paper is whether the proposed estimators are in any sense optimal. In the case of the direct effect, the $1/\sqrt{n}$ rate of convergence is clearly optimal; however, it would be interesting to investigate whether any tractable results on efficiency are available in our setting. Meanwhile, in the case of the indirect effect, the optimal rate of convergence itself remains open. Our proposed PC balancing estimator achieves a $\sqrt{\rho_n}$-rate of convergence, which intuitively appears to be a reasonably strong rate for this task. For purpose of benchmarking, consider estimation of treatment effects in a stochastic block model.
Fig 6: Comparison of a histogram of $\tau_{\text{IND}}^{\text{PC}}$ across $N = 1000$ simulations, and the Gaussian limit predicted by Theorem 6.

with $K_n = \rho_n^{-1}$ non-interacting blocks. Then, any simple block-level randomized algorithm could at best hope for a $1/\sqrt{K_n} = \sqrt{\rho_n}$ rate of convergence, whereas our PC balancing estimator can achieve this rate using unit-level randomization alone. Developing formal lower bounds for this problem, however, would of course be of considerable interest.

Another interesting direction for future work is in understanding the generality of our results, i.e., under what conditions we can plausibly expect estimators of the direct effect under network interference to achieve a $1/\sqrt{n}$-rate of convergence. Here, we started with a specific generative model, including anonymous interactions and a graphon model for the exposure graph; however, it’s plausible to us that a similar result would hold under more generality. Lovász and Szegedy (2006) show that a graphon limit arises naturally by considering any sequence of dense graphs $E^n$ with the property that, for any fixed graph $F$, the density of copies of $F$ in $E^n$ tends to a limit. Is it similarly possible to devise regularity assumptions on a sequence of exposure graphs $E^n$ and potential outcome functions $\{f_i(\cdot)\}_{i=1}^n$ under which the behavior of estimators for the direct effect is accurately predicted by graphon modeling?

**Acknowledgment.** We are grateful for helpful discussions with Guillaume Basse, Emmanuel Candès, Peng Ding, Trevor Hastie, Avi Feller, Betsy Ogburn, Fredrick Sävje and seminar participants at a number of venues. This work was partially supported by NSF grant DMS–1916163.

**SUPPLEMENTARY MATERIAL**

**Appendices**

We provide details about our simulation study and complete proofs for the results in the main text. Code to reproduce the experiments is available from https://github.com/lsn235711/random-graph-interference.

**REFERENCES**

ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics* 48 1452–1474.

ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis* 11 581–598.

ARONOW, P. M. and SAMII, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics* 11 1912–1947.
Athey, S., Eckles, D. and Imbens, G. W. (2018). Exact p-values for network interference. *Journal of the American Statistical Association* **113** 230–240.

Athreya, A., Fishkind, D. E., Tang, M., Priebel, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lysinski, V. and Qin, Y. (2017). Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research* **18** 8393–8484.

Baird, S., Bohren, J. A., McIntosh, C. and Ozler, B. (2018). Optimal design of experiments in the presence of interference. *Review of Economics and Statistics* **100** 844–860.

Basse, G. and Feller, A. (2018). Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association* **113** 41–55.

Basse, G. W., Feller, A. and Toulis, P. (2019). Randomization tests of causal effects under interference. *Biometrika* **106** 487–494.

Borgs, C., Chayes, J., Cohn, H. and Zhao, Y. (2019). An Lp theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions. *Transactions of the American Mathematical Society* **372** 3019–3062.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics* **43** 177–214.

Davis, C. and Kahane, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis* **7** 1–46.

Duflo, E., Greenstone, M., Pande, R. and Ryan, N. (2013). Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India. *The Quarterly Journal of Economics* **128** 1499–1545.

Gao, C., Lu, Y. and Zhou, H. H. (2015). Rate-optimal graphon estimation. *The Annals of Statistics* **43** 2624–2652.

Halloran, M. E. and Struchiner, C. J. (1995). Causal inference in infectious diseases. *Epidemiology* **142**–151.

Hu, Y., Li, S. and Wager, S. (2021). Average Treatment Effects in the Presence of Interference. *arXiv preprint arXiv:2104.03802*.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association* **103** 832–842.

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

Imbens, G. W. and Rubin, D. B. (2015). Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.

Leung, M. P. (2020). Treatment and spillover effects under network interference. *Review of Economics and Statistics* forthcoming.

Lóvász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B* **96** 933–957.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal* **16** S1–S23.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* **10** 1–51.

Parise, F. and Ozdaglar, A. (2019). Graphon games. In *Proceedings of the 2019 ACM Conference on Economics and Computation* 457–458.

Ross, N. (2011). Fundamentals of Stein’s method. *Probability Surveys* **8** 210–293.

Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability* **18** 1–9.

Sávio, F., Aronow, P. M. and Hudgens, M. G. (2021). Average treatment effects in the presence of unknown interference. *The Annals of Statistics* **49** 673–701.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **1135**–1151.

Tchetgen Tchetgen, E. J. and VanderWeele, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research* **21** 55–75.

VanderWeele, T. J. and Tchetgen, E. J. T. (2011). Effect partitioning under interference in two-stage randomized vaccine trials. *Statistics & Probability Letters* **81** 861–869.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge University Press.

Yu, Y., Wang, T. and Samworth, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323.

Zhang, Y., Levina, E. and Zhu, J. (2017). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika* **104** 771–783.