A/B TESTING UNDER IDENTITY FRAGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized online experimentation is a key cornerstone of the online world. The infrastructure enabling such methodologies is critically dependent on user identification. However, nowadays consumers routinely interact with online businesses across multiple devices which are often recorded with different identifiers for the same consumer. The inability to match different device identities across consumers leads to an incorrect estimation of various causal effects. Moreover, without strong assumptions about the device-user graph, the causal effects are not identifiable. In this paper, we consider the task of estimating global treatment effects (GATE) from a fragmented view of exposures and outcomes. Our experiments validate our theoretical analysis, and estimators obtained through our procedure are shown be superior to standard estimators, with a lower bias and increased robustness.

1 INTRODUCTION

A/B testing has become indispensable to online businesses for improving user experience and driving up revenue. The infrastructure which enables this is critically dependent on identifiers, such as cookies or mobile device IDs, traditionally used by websites and apps to track users' browsing behavior and provide personalized content and ads. However, the assumption about the availability of identifiers has become more and more tenuous as users increasingly rely on multiple devices. This means that a customer's effective persona as seen by the advertiser is broken into multiple units – a phenomenon known as 'identity fragmentation'(Coey & Bailey, 2016; Lin & Misra, 2021). Further, the use of third-party identifiers is increasingly being curbed, due to privacy concerns, by both governmental and non-governmental entities, through legislation such as the GDPR [1] and through the deprecation of third-party cookies and advertising identifiers such as the Android Advertising ID (AAID) and the Identifier for Advertisers (IDFA).

Lack of *identifiable information across devices creates a fundamental issue in A/B testing, as the users' exposure to treatment is not fully known in this setting*. Consider the case of a business exploring whether a certain advertisement produces a higher click-through rate. Under the standard A/B testing protocol, a random subset of users will be shown the new ad (B), and the outcome recorded. By comparing the outcomes for these users against the set of users who received ad A, one can estimate the relative change caused in the click-through rate by ad B. For a user who visits using different devices, for instance a smartphone and a tablet, the unique identifier (say IDFA), allows the server to consistently show the user only ad B. However, without identifiers, one cannot be certain of whether the current device should be in the treatment group or the control group. This happens because, while the treatment is administered at device level, the outcomes are dependent on user-level treatments. Thus, the outcome as observed for a device can potentially be affected by the treatment on other devices. This constitutes a *violation of the stable unit treatment – SUTVA assumption (Rubin, 1980) – which standard A/B testing relies upon*.

This phenomenon of treatments to a unit affecting outcomes for other units has been studied in causal literature (Hudgens & Halloran, 2008; LeSage & Pace, 2009) under the name of interference. It is also known as spillover, due to treatment exposure 'spilling over' from one unit to another. However, most methods involving spillover, assume strong restrictions on the structure of spillover (Ogburn et al., 2017; Leung, 2020). The deprecation of identifiers *introduces a new scenario, requiring the estimation of treatment effects from an uncertain interference structure.* This problem setting involves new assumptions compared to prior work. Notably, in addition to the assumption that unit/device

[1] `https://gdpr-info.eu/`

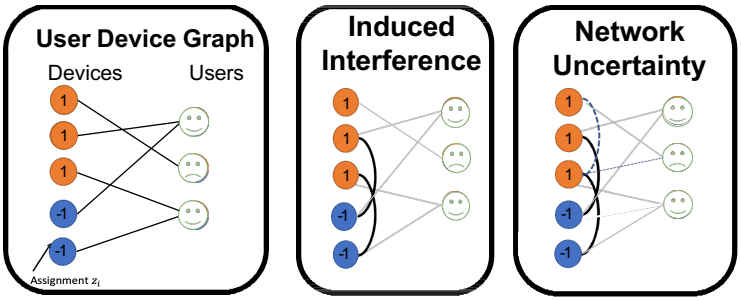


Figure 1: The user device graph presents the connections between the set of users and devices (Left). Treatments $Z_i \in \{-1, 1\}$ applied on a device, exposes the user of the device to the corresponding experience or ad. The outcomes depend on the total exposure a user has had to the treatment. As such the outcome at a device unit i now depends on the assignment of other devices j, which induces an interference graph between the devices (Middle). Under uncertain information the induced interference graph has potentially extra (dashed) or fewer edges (Right).

level outcomes are affected by treatments at other units/devices with the same user and not by those of other users, *an assumption can reasonably be made concerning the partial information about the device-user pairings, represented by a structure called 'the device graph'*. Partial information about the device graph be obtained, for instance, from devices with enabled cookies, from geolocation based on IP addresses or from an identity linking model (Sinha et al., 2014; Saha Roy et al., 2015).

In this work, we explore the problem of estimating the *global average treatment effect* (GATE) in the identify fragmentation setting, *under the assumption that interference comes only from devices that share the same user and that, for each user, a superset of their devices is known.* We formalize this problem as treatment effect estimation with interference, where the interference structure is based on the 'device neighborhood' i.e. the set of devices which share a user. We argue that the GATE is identifiable under reasonable assumptions. Finally, we propose a new VAE-based procedure that results in estimators that are superior to existing ones, as demonstrated through extensive experiments on both simulated and real data.

2 Related Work

2.1 Network Interference

Network interference is a well studied topic in causal inference literature, with a variety of methods proposed for the problem. Existing works in in this area incorporate various sets of assumptions to provide an estimate of treatment effects. A common approach is the exposure mapping framework which allows defines a degree of "belonging" of a unit to either the treatment or control group (Aronow et al., 2017; Auerbach & Tabord-Meehan, 2021; Li et al., 2021; Viviano, 2020). A common assumption is that the network effect is linear with respect to a known functional of the neighbour treatments(Basse & Airoldi, 2018; Cai et al., 2015; Chin, 2019; Gui et al., 2015; Toulis & Kao, 2013; Eckles et al., 2017; Sussman & Airoldi, 2017). A limitation of these approaches is that they require complete knowledge of the network structure. Similar to these proposals, our approach also relies on imposing an exposure-based structure to the form of interference, however we can also handle GLM-like outcomes as well an incomplete knowledge of the network.

Treatment effect estimation with unknown network interference has also been studied with the seminal work of Hudgens & Halloran (2008). The key insight behind these works is that if the network can be broken into clusters, then one can perform treatment effect estimation without the full knowledge of the interference structure withing the clusters. Other works such as Auerbach & Tabord-Meehan (2021); Bhattacharya et al. (2020); Liu & Hudgens (2014); Tchetgen & VanderWeele (2012); VanderWeele et al. (2014) have extended this idea further. Often the bias of these estimators depends on the the number of edges between the clusters, which has led to optimization-based methods for constructing clusters (Eckles et al., 2017; Gui et al., 2015). However, this still requires information about the clusters, and is not applicable if multiple clusters of the required type do not

exist. Finally, there are methods, which under restrictive assumptions, use SUTVA based estimates for one-sided hypothesis tests for treatment effect under interference (Choi, 2014; Athey & Wager, 2019; Lazzati, 2015).

Estimation without any side information: Recently, some methods have been proposed based on multiple measurements which can address the issue of interference(Shankar et al., 2023b; Cortez et al., 2022; Yu et al., 2022) without any further knowledge. However, such methods assume stationarity i.e. the outcomes do not vary between the trials. This simplifies GATE estimation by providing access to both the factual and counterfactual outcome. However, such a model is unrealistic for our motivating use case of continuous optimization. Furthermore, in the more general settings, conducting multiple trials can be difficult, if not impossible, in itself (Shankar et al., 2023a). As such, we aim to develop a method which can work with only a single trial and/or observational data from an existing test.

2.2 Estimation with Noisy Data

Parameter estimation with measurement noise is a well studied problem in causal inference (Wickens, 1972; Frost, 1979). Many methods and heuristics have been proposed for estimation of treatment effect (Carroll et al., 2006; Schennach, 2016; Ogburn & Vanderweele, 2013; Lockwood & McCaffrey, 2016). Yi et al. (2021) provides an overview of recent literature on the bias introduced by measurement error on causal estimation. Earlier works have focused on qualitative analysis by encoding assumptions of the error mechanism into a causal graph Hernán & Robins (2021), outcome Shu & Yi (2019), confounders Pearl (2012); Miles et al. (2018) and mediators Valeri & Vanderweele (2014).

Noisy covariates or proxy variables are not generally sufficient to identify causal effects (Kuroki & Pearl, 2014). As such works such as Kuroki & Pearl (2014); Miao et al. (2018); Shpitser et al. (2021); Dukes et al. (2021); Ying et al. (2021); Guo et al. (2022) have focused on identifying criteria for treatment effect estimation with noisy measurements with confounding variables.

Methods based on assuming knowledge of the error model are also common (Gustafson, 2003; Shpitser et al., 2021; Fang et al., 2023). Consequently, other methods for estimating causal effects also exist relying upon additional information such as repeated measurements (Shankar et al., 2023b; Cortez et al., 2022), instrumental variables (Zhu et al., 2022; Tchetgen et al., 2020) or a gold standard sample of measurements (Shankar et al., 2023a). A few works have also tried to study causal inference with measurement errors and no side information Miles et al. (2018); Pöllänen & Marttinen (2023). Other works have focused on partial identification of treatment effects (Zhao et al., 2017; Yadlowsky et al., 2018; Zhang & Bareinboim, 2021; Yin et al., 2021; Guo et al., 2022), sensitivity analysis (Imbens, 2003; Veitch & Zaveri, 2020; Dorie et al., 2016). Our work differs from these lines of work, as they usually focus on noisy measurements of unknown confounders or covariates, whereas our focus is on unknown network interference.

3 Notation

We are given a population of n devices. Let $\boldsymbol{Z}$ be the treatment assignment vector of the entire population and let $\mathcal{Z}$ denote the treatments' space, e.g., for binary treatments $\mathcal{Z} = \{-1, 1\}^n$. We use the Neyman potential outcome framework (Neyman, 1923; Rubin, 1974), and denote by $Y_i(\boldsymbol{z})$ the potential outcome for each $\boldsymbol{z} \in \mathcal{Z}$. We can make observations at only the device level, these observations are denoted as Y_i for device i. Note that the devices might have a common user, as presented in Figure 1. We assume that the outcome is determined by the user action, and hence the potential outcome at a device i need not depend only on its own treatment assignment but also other treatments allocated to the user's devices. This is a violation of the SUTVA assumption (Cox, 1958; Hudgens & Halloran, 2008); and is commonly called interference or spillover.

The user-device graph induces a dependence between device level outcomes. This dependence can also be represented as a device-level graph (Figure 1(Middle)), where each node represents a device and the presence of an edge indicates a common user between the device pair. The underlying graph is given by its adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, with $A_{ij} = 1$ only if an edge exists between devices i and j, and by convention $A_{ii} = 1$. Let $\mathcal{N}_i(\boldsymbol{A}) = \{j : A_{ij} = 1\}$ be the set of *neighbors* of device i. Since we assume the underlying graph is fixed, we will use $\mathcal{N}_i(\boldsymbol{A})$ and $\mathcal{N}_i$ interchangeably. We assume that the outcomes depend on the treatments received by a user (i.e. SUTVA holds at the user level). This implies that the interference at a device is limited to its neighbours in the graph.

User Level SUTVA: $\forall\ \boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}$ s.t. $z_i = z_i'$ and $z_j = z_j'\ \ \forall j \in \mathcal{N}_i$: $Y_i(\boldsymbol{z}) = Y_i(\boldsymbol{z}')$. **(A1)**

We will assume that the experimental design is a randomized Bernoulli design i.e. each device i gets allotted the treatment $z_i = 1$ independently with probability $p \in (0, 1)$. This is analogous to the standard randomization and positivity assumption in causal inference, and is equivalent if one assumes the exposure map $Y_i(\boldsymbol{z})$ only depends on z_i.

The desired causal effect is the mean difference between the outcomes when $\boldsymbol{z} = \vec{1}\, i.e.\, z_i = 1\, \forall i$ and when $\boldsymbol{z} = \vec{0}\, i.e.\, z_i = -1\, \forall i$. Under the aforementioned notations, this causal effect is given by:

$$\tau(\vec{1}, \vec{0}) = \frac{1}{n}\sum_{i=1}^{n} Y_i(\vec{1}) - \frac{1}{n}\sum_{i=1}^{n} Y_i(\vec{0}) \tag{1}$$

If the true graph $\boldsymbol{A}$ is known, under certain assumptions one can estimate the above treatment effect (Hudgens & Halloran, 2008; Halloran & Hudgens, 2016). However, in our problem setting, knowledge of the true graph would imply knowing which devices belong to the same user. As such we cannot assume, that $\boldsymbol{A}$ is known. However, we have access to some information about $\boldsymbol{A}$. In our use case of online experimentation, this information can come from those devices where the user has given cookie permissions, or from covariate information like geography or IP addresses, or from some existing model user for identity linking (Sinha et al., 2014).

Finally, we assume access to a model $\mathcal{M}$ which provides information on $\boldsymbol{A}$. Specifically, we assume that the $\mathcal{M}$ can be queried for any device i to get a predicted (or assumed) neighbours of a device (see Figure 1 (Right)). We will denote this neighbourhood by $\mathcal{M}(i)$.

Our primary focus revolves around estimating the Generalized Average Treatment Effect (GATE) under the previously outlined scenario, where there exists a degree of uncertainty concerning the network structure. Before we delve further into the method we provide a brief explanation of commonly used estimators and their problems for our problem setting.

Inverse Propensity/Horvitz-Thompson Estimate If the graph is known and when all treatment decisions are iid Bernoulli variables with probability p: one can use the classic Horvitz Thompson estimator as follows:

$$\frac{1}{n}\sum_i Y_i \left(\frac{\prod_{j\in\mathcal{N}_i} z_j}{\prod_{j\in\mathcal{N}_i} p} - \frac{\prod_{j\in\mathcal{N}_i}(1-z_j)}{\prod_{j\in\mathcal{N}_i}(1-p)} \right) = \frac{1}{n}\sum_i Y_i \left(\prod_{j\in\mathcal{N}_i} \frac{z_j}{p} - \prod_{j\in\mathcal{N}_i} \frac{(1-z_j)}{(1-p)} \right)$$

This inverse propensity estimators (and its derivatives) do not require any further assumption other than randomization and positivity to be unbiased. However, on inspection, one can see that this estimator ignores any units for which all neighbours are not in control or treatment groups. This results in extremely high variance, as most data samples are ignored. Moreover, if the number of neighbours is large, then this estimate may not even have a meaning, as there may not exist units for which all the neighbours are in control or treatment groups. This is particularly troublesome for our application as uncertainty in the graph means accounting for more possible units which interfere with a given unit, and including such units adds to the estimation issue of HT-estimators.

SUTVA Estimate The SUTVA estimate is given by

$$\hat{\tau}_{SUTVA} = \bar{Y}^{1} - \bar{Y}^{-1} = \frac{\sum Y_i \mathbb{I}[Z_i = 1]}{\sum \mathbb{I}[Z_i = 1]} - \frac{\sum Y_i \mathbb{I}[Z_i = -1]}{\sum \mathbb{I}[Z_i = -1]}$$

where $\bar{Y}^{-1/1}$ are the average of observed outcomes for units where $Z_i = -1/1$ respectively. Since it is the difference in means of control and treatment groups, it is also called the difference in mean/ DM estimator. This estimator while quite efficient and practical, requires the SUTVA assumption to be unbiased. As such these estimators can be misleading when it comes to our scenario.

4 METHOD

4.1 MODEL AND ASSUMPTIONS

Randomized experiments with interference (even with neighbourhood interference) can be difficult to analyze since the number of potential outcome functions grows exponentially: $2^{\mathcal{N}_i}$ for unit i; unlike the SUTVA case where one has only two outcomes. As such the literature around network interference restricts the space of potential outcome functions in order to do meaningful inference. One common approach is the exposure function (or exposure mapping) approach. Under this model one uses exposure variables which are functions from the discrete combinatorial space $\{-1, 1\}^{\mathcal{N}_i} \to \mathbb{R}^d$. One posits that the outcome Y_i depends on the treatment $\boldsymbol{z}$ only via the exposure variable e_i (Hudgens & Halloran, 2008; Aral & Walker, 2011; Aronow et al., 2017; Brennan et al., 2022). We will abuse notation, and often use e_i instead of the functional notation $e_i(\boldsymbol{z})$.

We too consider an exposure model; specifically we assume an outcome model of the form

$$Y_i(\boldsymbol{z}, x_i) = \underbrace{\mu_Y(\boldsymbol{z}, x_i)}_{\mathbb{E}[Y|Z=\boldsymbol{z}, X_i=x_i]} + \epsilon = c_0(x_i) + c_1(x_i)z_i + g(w(x_i)^T e_i(\boldsymbol{z}, x_i)) + \epsilon \qquad \textbf{(A2)}$$

where ϵ is mean zero noise, and x_i are the covariates at unit i. Assumption **A2** as stated is very generic, since the exposure function itself can be arbitrary. For meaningful inference, one often invokes a specific parametric form for the exposure function. A common example is an exposure represented as the (weighted) proportion of neighboring units that have received treatment (Eckles et al., 2017; Toulis & Kao, 2013). Alternatively, it could involve the count of neighboring units that have undergone treatment (Ugander et al., 2013). We will assume an additive vector exposure function along with some other standard assumptions (stated below) from treatment effect literature (Pearl, 2009).

$$\text{Additive Exposure: } e_i = \sum_{j \in \mathcal{N}_i} \phi(z_j, X_i) \qquad \textbf{(A3)}$$

$$\text{Network Ignorability: } Y(\boldsymbol{z}) \perp\!\!\!\perp \boldsymbol{Z}\ \forall \boldsymbol{z} \qquad \textbf{(A4)}$$

$$\text{Positivity: } P(\boldsymbol{z}|\boldsymbol{X}) > 0\ \forall \boldsymbol{z} \qquad \textbf{(A5)}$$

$$\text{Consistency: } Y_i = Y_i(\boldsymbol{z}) \text{ if } \boldsymbol{Z} = \boldsymbol{z} \qquad \textbf{(A6)}$$

$$\text{Neighbourhood Superset: } \mathcal{M}(i) \supseteq \mathcal{N}_i \qquad \textbf{(A7)}$$

Since ϕ in Assumption (**A3**) depends on the individual covariates, this assumption supports unit-level observed heterogeneity. We can also include the covariates x_j of the neighbouring units as well in ϕ but ignore this for simplicity. Further ϕ can be a vector function instead of scalar, and so **A3** can support all set function of neighbourhood treatments (Braun & Griebel, 2009). Moreover it also supports other common assumptions such as those in (Toulis & Kao, 2013; Eckles et al., 2017; Pouget-Abadie et al., 2019)

Remark 1. ***A7*** *can seem to be a strong assumption. However, in many applications, it is not difficult to satisfy this assumption. As a simple example, consider all devices which share a geographic location, with a given device i. This is very likely to be a superset of all devices that share a user with i. Furthermore, in practice, device-linking methods are used to identify neighbours based on confidence scores. These methods can usually be adapted to obtain a superset of neighbours with high probability (by including even low confidence nodes as neighbours).*

4.2 MODEL TRAINING

We propose using a latent variable model to infer the treatment effect. The dependence between various variables is depicted in Figure 2. We denote by E the true exposure which is the key latent variable of the model. $\tilde{E}$ is the exposure as implied by $\mathcal{M}$, which is our uncertain representation of the underlying device graph. The key difference between this and a standard exposure based causal model, is that in the latter the true exposure E is observed whereas in our model it is unobserved. Instead of E we observe the noise corrupted value $\tilde{E}$.

Remark 2. *Note that the true exposure E depends on the actual neighbourhood $\mathcal{N}_i$, while the observed exposure $\tilde{E}$ depends on the assumed neighbourhoods $\mathcal{M}(i)$.*

Fundamentally, this is a discrete problem as Z is a binary assignment of treatments at individual devices. However, since training such models is computationally intensive , we use a variational autoencoder (VAE) (Kingma & Welling, 2013; Kingma et al., 2019) based approximate training. In the appendix we argue why this procedure is analogous to the learning method suggested in Schennach & Hu (2013).

We posit a generative model for the joint distribution $p_\theta(\tilde{E}, E, Y|X, Z)$ which factorizes as $p_\theta(Y|E, X)p(\tilde{E}|E)p(E|Z)$. For the outcome distribution Y we posit a GLM style model which parameterizes $\mathbb{E}[Y|Z = z, X = x]$ from **A2** in terms of a neural network i.e. we use a neural network for each of the function c_0, c_1, g, w in **A2**. For the $p(\tilde{E}|E)$ we use a Gaussian model. If $|\mathcal{M}(i)| >> \mathcal{N}_i$, by law of large numbers this is a good approximation for the error. Finally $p(Z|X)$ is just the allocation mechanism which is exactly known to us as the experimenter.

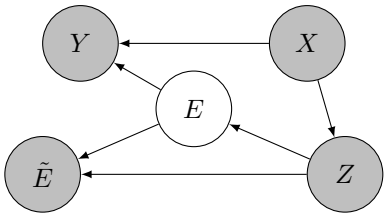


Figure 2: Graphical model depicting relationships between different variables for our model. Observed variables $\tilde{E}$ (noisy exposure), Y (effect/outcome), and X (covariates), Z (treatment allocation) are shaded to distinguish them from the hidden variable E (true treatment).

To use VAE style learning one needs to specify a posterior q_ϕ for the latent variable. For this we use a Gaussian variational approximation with both mean and variance parameterized. Specifically we use a q of the form $N(e|\mu_q(\tilde{e}, x, y; \phi), \sigma_q(\tilde{e}, x, y; \phi))$. As our objective function, we use the K-sample importance weighted ELBO $\mathcal{L}_K$ Burda et al. (2016), which is a lower bound for the conditional log-likelihood $p_\theta(x, y|z)$:

$$\mathcal{L}_K = \sum_{i=1}^{N} \mathbb{E}\left[\log \frac{1}{K} \sum_{j=1}^{K} w_{i,j}\right] \leq \sum_{i=1}^{N} \log \mathbb{E}\left[\frac{1}{K} \sum_{j=1}^{K} w_{i,j}\right] = \log p_\theta \tag{2}$$

where $w_{i,j} = p_\theta(\tilde{e}_i^*, z_{i,j}, x_i, y_i)/q_\phi(e_{i,j}|\tilde{e}_i, x_i, y_i)$ are importance weights, and the expectation is respect to q_ϕ. To reduce training variance we use the recent DReG estimator (Tucker et al., 2018). Once the model p_θ has been trained, one can obtain estimates of the mean outcomes $\mu_Y(\boldsymbol{z}, x_i)$ using $p_\theta(Y|E, X)$. By plugging the estimated outcomes into Equation 1, we get our estimate $\hat{\tau}$ [2].

Remark 3. *While the probability distribution can be arbitrarily parameterized with neural networks, all the neural networks used in our experiments, are MLPs with one hidden layer and ReLU activation.*

4.3 IDENTIFIABILITY

A key concern in causal inference, is the identifiability of the desired estimand, as otherwise there is no justification for the estimated value to correspond to the ground truth. Next, we discuss the identifiability of the treatment effect in the aforementioned scenario. The identifiability of treatment effect in our model is related to results in Schennach & Hu (2013). We summarize the crux of the argument below, while deferring the details to Appendix A

Proposition 1. *Under Assumptions **A1-7** and certain technical conditions on the function μ_Y, the conditional mean function $\mathbb{E}[Y|Z = z, X = x] = \mu_Y(x, z)$ is identifiable.*

Under **A2,4-6**, the problem of treatment effect estimation becomes a model fitting problem. Specifically, if the exposures e_i are known, one can conduct a regression of the observed outcomes Y_i on the exposures e_i and covariates X_i to estimate the population-level mean potential outcomes functions, denoted as μ_Y. Once we estimate the mean potential outcomes, we can obtain the treatment effect τ by plugging in these estimates into Equation 1.

When the graph A is exactly known, one can compute the exposures e_i using Assumption **A3**. However, since in our problem, the graph is unknown, obtaining e_i is not possible. To address this obstacle, we reframe the inference problem in our scenario as a regression with a measurement

[2]Refer to Appendix for more details

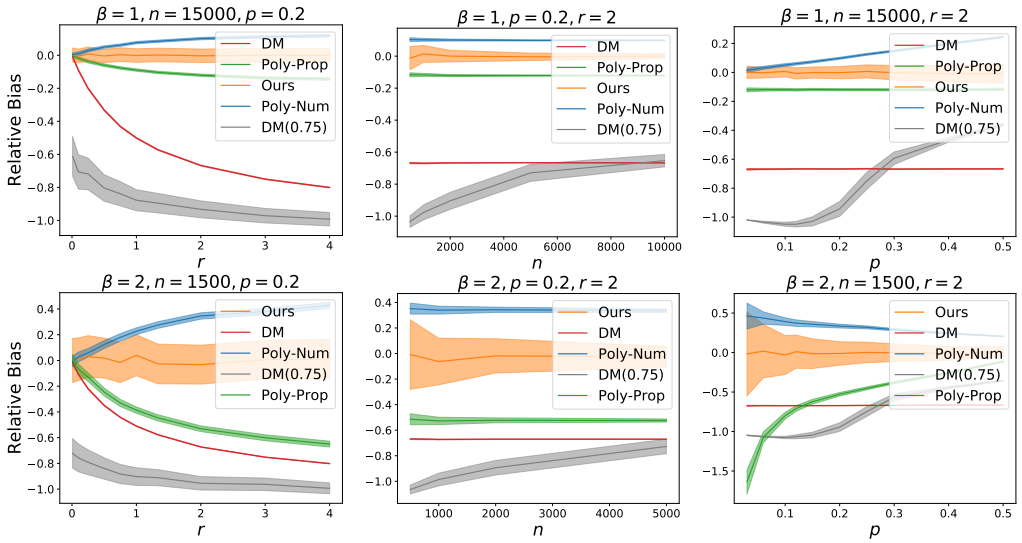
error problem. Observe that the exposure e_i under the assumed graph $\mathcal{M}$ is given by $e_i(\mathcal{M}) = \sum_{j \in \mathcal{M}(i)} \phi(z_j, X_i)$. Due to **A7** $e_i(\mathcal{M})$ can be decomposed as $e_i(\mathcal{N}_i) + \Delta e_i$, where Δe_i is an independent error term. Thus we can use $e_i(\mathcal{M})$ as noisy estimates of $e_i(\mathcal{N}_i)$.

Next, we argue the identifiability of the above regression task. Schennach & Hu (2013) provide conditions under which models of the form:

$$Y = \mu_Y(E) + \Delta Y; \quad \tilde{E} = E + \Delta E \;\; \Delta E \perp\!\!\!\perp E$$

can be identified from only the joint observations of $Y, \tilde{E}$. We show that the under assumptions **A1-6**, the conditions required for the identifiability results in Schennach & Hu (2013) are satisfied, thus making our model identifiable [3]. A detailed discussion is provided in the Appendix.

Remark 4. *This result does not apply when $\mathcal{M}(i) \subset \mathcal{N}_i$ because then the error term $\Delta\epsilon_i = \epsilon_i(\mathcal{M}) - \epsilon_i(\mathcal{N}_i)$ is no longer independent of the true exposure $\epsilon_i(\mathcal{N}_i)$. In that case, the our approach becomes equivalent to regression with endogenous covariate error, which requires additional information ().*



(a) Varying direct:indirect effects (r) (b) Varying population size (n) (c) Varying treatment budget (p)

Figure 3: Plots visualizing the performance of various GATE estimators under Bernoulli design on Erdos-Renyi networks for both linear and quadratic potential outcomes models. The lines represent the empirical relative bias i.e. $\frac{\hat{\tau}-\tau}{\tau}$ of the estimators across different settings, with the shaded width corresponding to the experimental standard error.

5 Experiments

5.1 Synthetic Graphs

In this section, we first experimentally demonstrate the validity of our approach by experimenting with synthetic data obtained from a model which satisfies our assumptions exactly. For this we experiment with synthetically generated Erdos-Renyi graphs to compare the performance of our estimator with other estimators. We simulate 100 different random graphs and run repeated experiments on this graph with random treatment assignments. We sample covariates X independently from a multivariate normal distribution and consider a polynomial family of outcome models. Specifically the outcomes are simulated from the following equation $Y_i(\boldsymbol{z}, X_i) = c_0(X_i) + g(w(X_i)^T \sum_{j \in \mathcal{N}_i} \phi_{i,j}(z_j)) + \epsilon$ where g is a polynomial function of order β and ϵ is mean 0 error. Similar to Cortez et al. (2022), we experiment with the linear $\beta = 1$ and quadratic $\beta = 2$ setting. For each experiment, we varied the treatment probability p, the size of the graphs n to assess the efficacy of estimation across different ranges of parameters and the strength of interference r. Following Cortez et al. (2022), the strength of

[3]The primary restriction is that g should not be of the form $g(z) = a + b\ln(\exp(cz) + d)$

interference is measured as the ratio of norms of the self-influence $\phi_{i,i}$ and average cross-influences $\phi_{i,j}$ i.e. $r = \frac{1}{n}\sum_i \frac{\sum_{j\in\mathcal{N}_i\setminus i}|\phi_{i,j}|}{|\phi_{i,i}||\mathcal{N}_i|}$

Baselines In our evaluation, we gauge the effectiveness of our proposed method by benchmarking it against commonly employed estimators such as polynomial regression (Poly), difference-in-means (DM) estimators. Since the polynomial regression model needs exact neighbourhoods, we use them in an oracle setting i.e. they have access to the true device graph. [4]

The results are presented in Figure 3. From the figure it is clear that our model produces unbiased estimates in this case. On the other hand, all other methods produce highly biased estimates. Note that in Figure 3a, when $r = 0$, there is no interference, and hence most estimators are unbiased. However, when interference increases these methods clearly show strong bias. Secondly, for a given interference strength, our method shows consistency in the form of decreasing variance with increasing number of nodes. Finally, the variance of our method reduces as the treatment probability p increases to 0.5.

5.2 AIRBNB SIMULATIONS

Next, we conduct simulation from a model designed from the AirBnB vacation rentals domain Li et al. (2022). The original model is a simulator for rental listings and their bookings for a two-sided marketplace. Contrary to the previous experiments, the outcomes here do not follow an explicit exposure mapping. We adapt this simulator for our purposes, replacing customers with devices and listings with users. The measured outcome Y_i is 1 iff there is a click on device i. A user watches ads on a randomly chosen subset of its devices, and chooses to click on the ad on only one device, leading to interference between outcomes. This simulation works uses a type matching model where if the device and person have the same type, the probability of watching an ad on that device is higher. The treatment scales the probability of seeing an ad by the parameter α. This is a good testbed for testing robustness of our model, since like in the real-world, exposure models are only our best approximations to the unknown and complex actual interference function. We perform simulations with protocol specified in Brennan et al. (2022) [5].

Baselines As baselines in this experiment, we use the SUTVA/DM estimator, an exposure model with oracle graph i.e. one where the exact graph is known (labelled Exp), and a Horvitz-Thompson estimator with oracle graph (labelled HT). The Exp model is same as the one used in Brennan et al. (2022), while the HT estimator is the one described in Section 3. The performance of different estimators is shown in Figure 4.

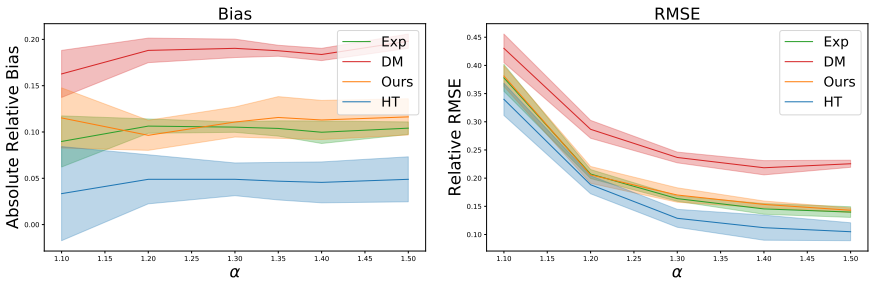


Figure 4: Visualization of performance of different GATE estimators on the airbnb simulator. The lines represent a) absolute relative bias $\left|\frac{\hat{\tau}-\tau}{\tau}\right|$ and b) relative RMSE of various algorithms as the indirect treatment effect α increases. Bands capture the standard deviation over 500 trials.

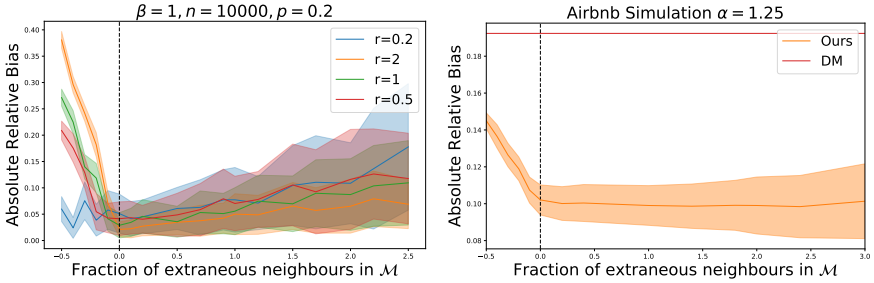
Since the exposure model can only partly model the actual outcomes, in this case, bias is not zero. On the other hand, the Oracle HT estimator (which makes no exposure assumptions) gives unbiased though higher variance estimates. The model is Oracle in using the exact interference graph. A different model is the Oracle Exposure (Exp) model which used the true graph to compute the

[4] Due to incorporating large neighbourhoods (with upto 100 extraneous nodes), Horwitz-Thompson type estimators failed to yield non-meaningful results in any trial.

[5] Details in Appendix

exposure. From the result it is also clear that our approach works as well as the Oracle Exposure model. Furthermore, even on the MSE metric our model performs comparably to the Exp model. These results suggest that our method is robust even when the true potential outcome does not obey the assumed exposure mapping.

5.3 Effect of Network Uncertainty



(a) Erdos-Renyi Networks (b) AirBnB Simulator

Figure 5: Impact of neighbourhood sizes on the absolute relative bias i.e. $|\frac{\hat{\tau}-\tau}{\tau}|$ GATE estimation. Negative fraction of neighbours indicate the case when $\mathcal{M}(i) \subset \mathcal{N}_i$ i.e. we missed pertinent neighbours. The bias tends to be high when gives small neighbourhoods, as they miss pertinent edges. As the neighbourhood sizes increase, the bias reduces, but the uncertainty widens.

Next we examine the impact of the neighborhood accuracy $\mathcal{M}(i)$ in estimation. We experiment with Erdos-Renyi graphs as well as with the AirBnB Model. For these experiments, we fix a single graph, and compute the treatment effect estimate from our method as we change the assumed neighbourhoods $\mathcal{M}(i)$. In Figure 5a, we preset the relative ratio between the estimated and true treatment effects as varying proportions of edges are either added or omitted by $\mathcal{M}(i)$. To maintain simplicity, we maintain uniform $\mathcal{M}(i)$ sizes across all nodes, employing the average number of missed or added edges as the metric along the x-axis. Figure 5b presents the same experiment within the context of the Airbnb simulator. We observe a similar trend in both experiments: when $\mathcal{M}(i) \supseteq \mathcal{N}_i$ holds true for all nodes i, our approach can offer an lower bias estimate of the treatment effect. Nonetheless, as the number of extraneous nodes within $\mathcal{M}(i)$ grows, so does the uncertainty in estimation. Conversely, if $\mathcal{M}(i)$ neglects a pertinent node, it may introduce greater bias into the estimation process. This manifests within our results, where the model predictions initially exhibit strong bias. However, as neighborhood sizes expand, bias diminishes while variance increases.

5.4 Application: Assessing Power Plant Emissions Controls

We use our approach to estimate the effect of pollution reduction technologies on ambient ozone levels. As ambient pollution is heavily influenced by spatially adjacent sources of pollution, adjusting for interference is important. DM estimators in this case often underestimate the impact in these scenarios. We work with a public dataset on 473 power generation facilities in USA used in Papadogeorgou et al. (2019). We use the DM, Poly and Exp estimators as baselines of which the latter two need exact neighbourhoods. For our method we will not use coordinate information for identifying neighbourhoods and instead uses groupings based on census divisions. The results (Figure: 6) show that our method provides comparable estimates to other oracle estimators.

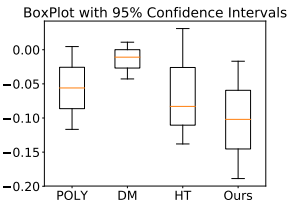


Figure 6: GATE on ambient ozone levels of adopting of SCR/SNCR technologies

6 Conclusion

Identity fragmentation is an increasingly relevant problem in online A/B testing. Our work provides a method to estimate GATE under a relaxed assumption of having knowledge only about the super-set of the identities that belong to the user. This relaxed assumption can be practically far more feasible than requiring the exact network. With both theoretical and experimental analysis, we established the efficacy of our estimator(s) under this assumption.

REFERENCES

Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management science*, 57(9):1623–1639, 2011.

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

Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2):37–51, 2019.

Eric Auerbach and Max Tabord-Meehan. The local approach to causal inference under network interference. *arXiv preprint arXiv:2105.03810*, 2021.

Guillaume W Basse and Edoardo M Airoldi. Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika*, 105(4):849–858, 2018.

Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 1028–1038. PMLR, 22–25 Jul 2020. URL `https://proceedings.mlr.press/v115/bhattacharya20a.html`.

Jürgen Braun and Michael Griebel. On a constructive proof of kolmogorov's superposition theorem. *Constructive approximation*, 30:653–675, 2009.

Jennifer Brennan, Vahab Mirrokni, and Jean Pouget-Abadie. Cluster randomized designs for one-sided bipartite experiments. *Advances in Neural Information Processing Systems*, 35:37962–37974, 2022.

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Jing Cai, Alain De Janvry, and Elisabeth Sadoulet. Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108, 2015.

Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006.

Alex Chin. Regression adjustments for estimating the global treatment effect in experiments with interference. *Journal of Causal Inference*, 7(2), 2019.

David Choi. Estimation of monotone treatment effects in network experiments. *Journal of the American Statistical Association*, 112(519):1147–1155, 2014.

Dominic Coey and Michael Bailey. People and cookies: Imperfect treatment assignment in online experiments. In *Proceedings of the 25th International Conference on World Wide Web*, WWW 16, 2016.

Mayleen Cortez, Matthew Eichhorn, and Christina Lee Yu. Graph agnostic estimators with staggered rollout designs under network interference. *Advances in Neural Information Processing Systems*, 2022.

David Roxbee Cox. Planning of experiments. 1958.

Vincent Dorie, Masataka Harada, Nicole Bohme Carnegie, and Jennifer Hill. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in Medicine*, 35(20): 3453–3470, 2016.

Oliver Dukes, Ilya Shpitser, and Eric J Tchetgen Tchetgen. Proximal mediation analysis. *arXiv preprint arXiv:2109.11904*, 2021.

Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1), 2017.

Paul Erdos, Alfred Renyi, et al. On the evolution of random graphs. *Publ. math. inst. hung. acad. sci*, 5(1):17–60, 1960.

Xiaoqiong Fang, Andy W Chen, and Derek S Young. Predictors with measurement error in mixtures of polynomial regressions. *Computational Statistics*, 38(1):373–401, 2023.

Peter A Frost. Proxy variables and specification bias. *The Review of Economics and Statistics*, pp. 323–325, 1979.

Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han. Network a/b testing: From sampling to estimation. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 399–409. International World Wide Web Conferences Steering Committee, 2015.

Wenshuo Guo, Mingzhang Yin, Yixin Wang, and Michael Jordan. Partial identification with noisy covariates: A robust optimization approach. 2022. URL `https://openreview.net/forum?id=-NVBxy0TdU`.

Paul Gustafson. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press, 2003.

M Elizabeth Halloran and Michael G Hudgens. Dependent happenings: a recent methodological review. *Current epidemiology reports*, 3(4):297–305, 2016.

Miquel A Hernán and James M Robins. *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC, 2021.

Michael G. Hudgens and M. Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

Ramesh Johari, Hannah Li, Inessa Liskovich, and Gabriel Y Weintraub. Experimental design in two-sided platforms: An analysis of bias. *Management Science*, 2022.

Jeffrey O Kephart and Steve R White. Directed-graph epidemiological models of computer viruses. In *Computation: the micro and the macro view*, pp. 71–102. World Scientific, 1992.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Natalia Lazzati. Treatment response with social interactions: Partial identification via monotone comparative statics. *Quantitative Economics*, 6(1):49–83, 2015.

James LeSage and Robert Kelley Pace. *Introduction to spatial econometrics*. Chapman and Hall/CRC, 2009.

Michael P Leung. Treatment and spillover effects under network interference. *Review of Economics and Statistics*, 102(2):368–380, 2020.

Hannah Li, Geng Zhao, Ramesh Johari, and Gabriel Y Weintraub. Interference, bias, and variance in two-sided marketplace experimentation: Guidance for platforms. In *Proceedings of the ACM Web Conference 2022*, pp. 182–192, 2022.

Wenrui Li, Daniel L Sussman, and Eric D Kolaczyk. Causal inference under network interference with noise. *arXiv preprint arXiv:2105.04518*, 2021.

Tesary Lin and Sanjog Misra. The identity fragmentation bias. *Available at SSRN 3507185*, 2021.

Lan Liu and Michael G. Hudgens. Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statistical Association*, 109(505):288–301, 2014. doi: 10.1080/01621459.2013.844698. URL https://doi.org/10.1080/01621459.2013.844698. PMID: 24659836.

JR Lockwood and Daniel F McCaffrey. Matching and weighting with functions of error-prone covariates for causal inference. *Journal of the American Statistical Association*, 111(516):1831–1839, 2016.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Caleb H Miles, Joel Schwartz, and Eric J Tchetgen Tchetgen. A class of semiparametric tests of treatment effect robust to confounder measurement error. *Statistics in Medicine*, 37(24):3403–3416, 2018.

Jerzy Neyman. On the Application of Probability Theory to Agricultural Experiments: Essay on Principles. *Statistical Science*, 5:465–80, 1923. Section 9 (translated in 1990).

Elizabeth L Ogburn and Tyler J Vanderweele. Bias attenuation results for nondifferentially mismeasured ordinal and coarsened confounders. *Biometrika*, 100(1):241–248, 2013.

Elizabeth L Ogburn, Oleg Sofrygin, Ivan Diaz, and Mark J Van der Laan. Causal inference for social network data. *arXiv preprint arXiv:1705.08527*, 2017.

Georgia Papadogeorgou, Christine Choirat, and Corwin M Zigler. Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. *Biostatistics*, 20(2):256–272, 2019.

Georgia Papadogeorgou, Kosuke Imai, Jason Lyall, and Fan Li. Causal inference with spatio-temporal data: Estimating the effects of airstrikes on insurgent violence in iraq. *arXiv preprint arXiv:2003.13555*, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. On measurement bias in causal inference. *arXiv preprint arXiv:1203.3504*, 2012.

Antti Pöllänen and Pekka Marttinen. Identifiable causal inference with noisy treatment and no side information. *arXiv preprint arXiv:2306.10614*, 2023.

Jean Pouget-Abadie, Kevin Aydin, Warren Schudy, Kay Brodersen, and Vahab Mirrokni. Variance reduction in bipartite experiments through correlation clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Donald B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980. ISSN 01621459. URL http://www.jstor.org/stable/2287653.

Rishiraj Saha Roy, Ritwik Sinha, Niyati Chhaya, and Shiv Saini. Probabilistic deduplication of anonymous web traffic. In *Proceedings of the 24th International Conference on World Wide Web*, WWW 15, 2015.

Susanne M Schennach. Recent advances in the measurement error literature. *Annual Review of Economics*, 8:341–377, 2016.

Susanne M Schennach and Yingyao Hu. Nonparametric identification and semiparametric estimation of classical measurement error models without side information. *Journal of the American Statistical Association*, 108(501):177–186, 2013.

Comandur Seshadhri, Tamara G Kolda, and Ali Pinar. Community structure and scale-free collections of erdos-renyi graphs. *Physical Review E*, 85(5):056109, 2012.

Shiv Shankar, Ritwik Sinha, Saayan Mitra, Moumita Sinha, and Madalina Fiterau. Direct inference of effect of treatment (diet) for a cookieless world. In *Proceedings of the The 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023a.

Shiv Shankar, Ritwik Sinha, Saayan Mitra, Viswanathan (Vishy) Swaminathan, Sridhar Mahadevan, and Moumita Sinha. Privacy aware experiments without cookies. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23. Association for Computing Machinery, 2023b.

Ilya Shpitser, Zach Wood-Doughty, and Eric J Tchetgen Tchetgen. The proximal id algorithm. *arXiv preprint arXiv:2108.06818*, 2021.

Di Shu and Grace Y Yi. Causal inference with measurement error in outcomes: Bias analysis and estimation methods. *Statistical methods in medical research*, 28(7):2049–2068, 2019.

Ritwik Sinha, Shiv Saini, and N Anadhavelu. Estimating the incremental effects of interactions for marketing attribution. In *2014 International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC2014)*, pp. 1–6. IEEE, 2014.

Daniel L Sussman and Edoardo M Airoldi. Elements of estimation theory for causal effects in the presence of network interference. *arXiv preprint arXiv:1702.03578*, 2017.

Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012. doi: 10.1177/0962280210386779. URL https://doi.org/10.1177/0962280210386779. PMID: 21068053.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Panos Toulis and Edward Kao. Estimation of causal peer influence effects. In *International Conference on Machine Learning*, pp. 1489–1497, 2013.

George Tucker, Dieterich Lawson, Shixiang Gu, and Chris J Maddison. Doubly reparameterized gradient estimators for monte carlo objectives. *arXiv preprint arXiv:1810.04152*, 2018.

Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 329–337. ACM, 2013.

Linda Valeri and Tyler J Vanderweele. The estimation of direct and indirect causal effects in the presence of misclassified binary mediator. *Biostatistics*, 15(3):498–512, 2014.

Tyler J. VanderWeele, Eric J. Tchetgen Tchetgen, and M. Elizabeth Halloran. Interference and sensitivity analysis. *Statist. Sci.*, 29(4):687–706, 11 2014. doi: 10.1214/14-STS479. URL https://doi.org/10.1214/14-STS479.

Victor Veitch and Anisha Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *arXiv preprint arXiv:2003.01747*, 2020.

Davide Viviano. Experimental design under network interference. *arXiv preprint arXiv:2003.08421*, 2020.

Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings.*, pp. 25–34. IEEE, 2003.

Michael R Wickens. A note on the use of proxy variables. *Econometrica: Journal of the Econometric Society*, pp. 759–761, 1972.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*, 2018.

Grace Y. Yi, Aurore Delaigle, and Paul Gustafson. *Handbook of Measurement Error Models*. CRC Press, 2021.