On Online Experimentation without Device Identifiers

Shiv Shankar [1] **Ritwik Sinha** [2] **Madalina Fiterau** [1]

Abstract

Measuring human feedback via randomized experimentation is a cornerstone of data-driven decision-making. The methodology used to estimate user preferences from their online behaviours is critically dependent on user identifiers. However, in today's digital landscape, consumers frequently interact with content across multiple devices, which are often recorded with different identifiers for the same consumer. The inability to match different device identities across consumers poses significant challenges for accurately estimating human preferences and other causal effects. Moreover, without strong assumptions about the device-user graph, the causal effects might not be identifiable. In this paper, we propose HIFIVE, a variational method to solve the problem of estimating global average treatment effects (GATE) from a fragmented view of exposures and outcomes. Experiments show that our estimator is superior to standard estimators, with a lower bias and greater robustness to network uncertainty.

1. Introduction

Using human feedback to align AI systems with the user's goals and preferences is an important problem, especially as these models are becoming widely used in critical areas including health [107] and education [109]. A commonly used workflow for obtaining human feedback involves presenting users with multiple versions of content (produced by a language model) and obtaining their preferences. For example, we have variations for the task, "a marketer would like to target a segment of customers, and uses generative AI to create images for the campaign". A principled method of learning about which outputs are preferred by users is to use randomized control trials, also known as A/B tests [8]. In a basic version of this procedure, two variants of a system or their outputs (e.g., versions of LLM generated completions), denoted as variant A and variant B, (sometimes also called control and treatment) are compared by randomly assigning them to the end-users and evaluating the metrics of concern on the two groups. For online businesses, A/B testing is crucial for evaluating users' preferences and experiences with the product [67, 2, 105].

The digital technologies that enable A/B testing are 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 [83, 67]. However, this assumption about the availability of identifiers has become more and more tenuous. Users have become increasingly reliant on multiple devices. At the same time, the use of third-party identifiers is being curbed, due to privacy concerns, by both governmental and non-governmental entities, through legislation such as the GDPR [29] and through the deprecation of third-party cookies and mobile advertising identifiers. This means that a customer's effective persona that can be observed online is broken into multiple units – a phenomenon known as 'identity fragmentation' [20, 52].

Lack of identifiable information across devices creates a fundamental issue in A/B testing, as the users' exposure to treatment is not controllable. Consider the case of a business exploring whether a new model produces content that is preferred by its users. Under the standard A/B testing protocol, a random subset of users will be shown the new content B, and their feedback recorded. By comparing the results for these users against the set of users who received content A, one can estimate the relative preferences of users for output B over A. 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 content 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 [69] – which causal inference from A/B tests relies upon*.

[1]College of Information and Computer Sciences, University of Massachusetts, USA [2]Adobe Research, USA. Correspondence to: Shiv Shankar <sshankar@umass.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

This phenomenon of treatments to a unit affecting outcomes for other units has been studied in causal literature [38, 46] 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 [57, 48]. The deprecation of identifiers *introduces a new scenario, requiring the estimation of treatment effects on an uncertain network structure.* This problem setting involves new assumptions compared to prior work. Notably, in addition to the assumption that unit/device 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 the 'device graph'*. Often, some information about the device graph can be obtained, for instance, from devices with login information, from geolocation based on IP addresses or from an identity linking model [87, 71].

Contribution In this work, we explore the problem of estimating the *global average treatment effect* (GATE) under identity fragmentation *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 uncertain network interference, where the interference graph is based on the 'device neighbourhood', i.e., the set of devices which share a user. Unlike other works on interference, *we do not assume any of the following: a) fully known network structure, b) linear outcomes or c) repeated measurements/multiple trials*. We propose a variational inference-based model called HIFIVE (**H**uman **I**nterest-estimate under **F**ragmented **I**dentities via **V**ariational **E**stimation) to estimate the GATE and show that the proposed model *is identifiable in this setting*. Through extensive experiments on both simulated and real data we show that our method is superior to other interference-aware methods while making weaker assumptions.

2. Related Work

Network Interference Network interference is a well studied topic in causal inference literature [10, 16, 18, 32, 92, 81]. Common approaches include assumptions about the interference neighbourhood [9, 94] or linear interference model [25, 88]. A limitation of these approaches is that they require complete knowledge of the network structure, while we consider *an incomplete knowledge of the network*. Recently, some methods based on multiple measurements have been proposed to address the issue of interference [83, 21, 112] without any further knowledge about its structure. However, such methods assume stationarity, i.e., the outcomes do not vary between the trials. This simplifies GATE estimation by implicitly 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 more general settings, conducting multiple trials can be difficult, if not impossible [82]. Thus, we aim to develop a *method which can work with only a single trial and/or observational data from an existing test*.

We summarize some common approaches, and how our method differs from them in Table 1. To the best of our knowledge, our method is the only one that can handle: *a) non-linearity in outcomes; b) works with un-structured graphs; c) without exact knowledge of the graph edges and d) without multiple trials and e) without side information*. A more detailed survey of the relevant interference literature is in the Appendix.

Table 1: Literature Summary. We list a few important works, criteria relevant to our work, and whether the criteria are satisfied ✓ or not ×. Our method is the only method which satisfies all criteria.

	General Graph	Uncertain Edges	Non-Linear Outcome	Single Trial
[38, 53]	×	✓	✓	✓
[113, 112]	✓	×	×	✓
[21, 83]	✓	✓	✓	×
[4, 72]	✓	×	✓	✓
[92, 25, 88]	✓	×	×	✓
HIFIVE	✓	✓	✓	✓

Estimation with Noisy Data Many methods and heuristics have been proposed for estimation of treatment effect [17, 73, 58, 54] with measurement noise in data. Yi et al. [110] provides an overview of recent literature on the bias introduced by measurement error on causal estimation. Many works have focused on qualitative analysis by encoding assumptions of the error mechanism into a causal graph [37, 80], outcome [86], confounders [63, 55] and mediators [95]. Methods based on knowledge of the error model are also common [34, 85, 30]. Existing proposals for estimating causal effects under noise rely upon additional information such as repeated measurements [83, 21, 79], instrumental variables [118, 91] or a gold standard sample of measurements [82]. While few works have also tried to study causal inference with measurement errors and no side information [55, 64], these works focus on noisy measurements of unknown confounders or covariates, *whereas our focus is on uncertain network interference*. Finally, some works have considered partial identification of treatment effects [116, 108, 115, 111, 33] and sensitivity analysis [39, 98, 24].
.

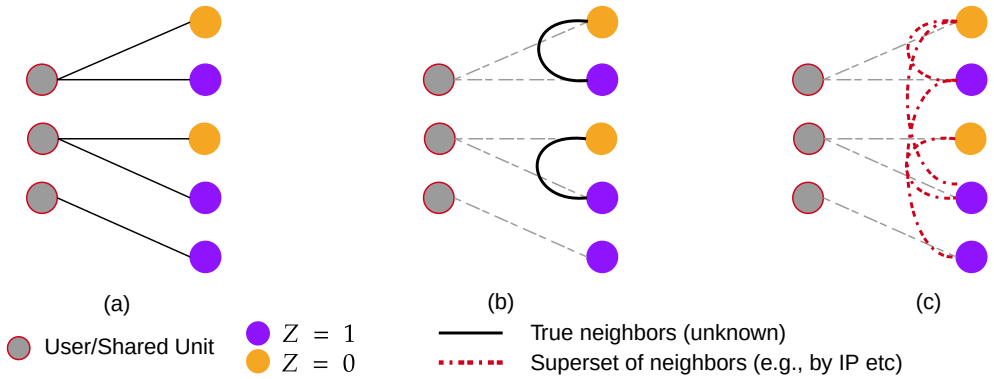


Figure 1: A bipartite graph (left) presents the connections between the set of users and devices. Treatments Z_i applied on a device expose the user of the device to the corresponding experience or algorithm etc. The outcomes depend on the total exposure to the treatment, hence the outcome at device i depends on the treatment at other device j, which induces an interference graph (Middle). Under link uncertainty the interference graph has potentially extra (dashed) edges (Right).

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 $\mathcal{Z} = \{0,1\}^n$. We use the Neyman potential outcome framework [56, 68], and denote by $Y_i(\boldsymbol{z})$ the potential outcome for each $\boldsymbol{z} \in \mathcal{Z}$. We can only make observations at device level, and denote these observations as Y_i for device i. Additionally we may have access to covariates X_i at the devices. 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 [23, 38] and is commonly called interference or spillover.

The user-device graph induces a dependence between device level outcomes. This dependence can be represented in an induced device-device graph (Figure 1, middle), where each node represents a device and the presence of an edge indicates a common user of the two devices. 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 *neighbours* of device i in the device-device graph. 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). Thus the interference is limited to a node's neighbours in the device-device graph. Formally this is equivalent to classic network neighbourhood interference assumption [38, 88], formally stated next, on the induced device-device graph.

Network Interference

$$\forall \boldsymbol{z}, \boldsymbol{z}' \text{ s.t. } z_i = z'_i \text{ and } z_j = z'_j \ \forall j \in \mathcal{N}_i : \quad Y_i(\boldsymbol{z}) = Y_i(\boldsymbol{z}'). \quad \textbf{(A0)}$$

We will consider randomized Bernoulli designs i.e., each device i gets allotted the treatment $z_i = 1$ independently with probability $p_i \in (0,1)$. This is natural and easy to implement, and satisfies standard randomization and positivity assumption in causal inference.

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 $\boldsymbol{z} = \vec{0}\, i.e.,\ z_i = 0\,\forall i$. Under the aforementioned notations, this causal effect is given by:

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

If the true graph $\boldsymbol{A}$ is known, under certain assumptions one can estimate the above treatment effect [38, 35]. 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. Instead we assume access to a model $\mathcal{M}$ which provides information on $\boldsymbol{A}$. Specifically, we assume that the model $\mathcal{M}$ can be queried for any device i to get predicted (or assumed) neighbours of a device (see Figure 1, right). We will denote this neighbourhood by $\mathcal{M}(i)$. Our method is agnostic to how $\mathcal{M}$ was formed, and so in this work we consider $\mathcal{M}$ as given. Often time, some information can be obtained by using meta-information such as IP, geo-locations or from users who have given permission for device linking. This provides a significant practical advantage over the prior methods that necessitate knowledge of the exact neighbourhood.

Our primary focus revolves around estimating the Global Average Treatment Effect (GATE) under the previously outlined scenario, where there exists a degree of uncertainty concerning the network structure. As such we want an approach which is agnostic to how $\mathcal{M}$ is obtained and robust to variations in it. Furthermore we would like to impose only constraints on $\mathcal{M}(i)$ that are easy to satisfy. Before we delve further into the method we provide a brief explanation of commonly-used estimators and why they are not suitable 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:

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

This inverse propensity estimator (and its derivatives) do not require any further assumption other than randomization and positivity. However, this estimator ignores any units for which all neighbours are not in control or treatment groups. If the number of neighbours is large, then this estimate may not be meaningful, 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 (or the DM estimate) is given by

$$\hat{\tau}_{\text{SUTVA}} = \bar{Y}^1 - \bar{Y}^0 = \frac{\sum Y_i \mathbb{I}[Z_i = 1]}{\sum \mathbb{I}[Z_i = 1]} - \frac{\sum Y_i \mathbb{I}[Z_i = 0]}{\sum \mathbb{I}[Z_i = 0]}$$

where $\bar{Y}^{0/1}$ are the average of observed outcomes for units where $Z_i = 0/1$ respectively. This estimator, while simple and practical, requires the SUTVA assumption, and hence can be misleading in our scenario.

4. HIFIVE

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. For meaningful inference, one often invokes an exposure mapping framework [38, 2, 4, 14]. Under this approach, one uses exposure variables e_i which are specific parametric functions mapping the discrete combinatorial space $\{0,1\}^{\mathcal{N}_i} \to \mathbb{R}^d$. One posits that the outcome Y_i depends on the treatment $\boldsymbol{z}$ only via the exposure variable e_i i.e. $Y_i = Y_i(e_i(\boldsymbol{z}))$. A common example is an exposure represented as the (weighted) proportion of neighbouring units that have received treatment [25, 92]. Alternatively, it could involve the count of neighbouring units that have undergone treatment [94]. We too consider an exposure model, but unlike most earlier works we allow for non-linearities in the model (**A1**). We will also assume that for each node i, the assumed neighbours $\mathcal{M}(i)$ are a superset of its true neighbours (**A2**). We shall call the nodes in the set $\mathcal{M}(i) - \mathcal{N}_i$ as extraneous, where '$-$' denotes set difference.

Exposure Assumptions

Exposure Model: $Y_i(\boldsymbol{z}, x_i) =$
$= \mathbb{E}[Y_i | \boldsymbol{Z} = \boldsymbol{z}, X_i = x_i] + \epsilon$
$= c_0(x_i) + c_1(x_i) z_i + g(w_i^T \sum_{j\in\mathcal{N}_i} \phi(z_j, X_i)) + \epsilon$ **(A1)**

Neighbourhood Superset: $\mathcal{M}(i) \supseteq \mathcal{N}_i$ **(A2)**

Here ϵ is mean zero noise, and x_i are the covariates at unit i. We will sometimes denote $\sum \phi(z, X)$ as just the exposure e_i. Since ϕ in 4.1 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 we supress this for simplicity.

Remark 1. Note that unlike most exposure models, we allow ϕ to be a vector function instead of scalar. Due to using the vector ϕ, (**A1**) can support all set functions of neighbourhood treatments [13, 75, 44]. This subsumes other common assumptions such as those used in [92, 25, 65]. Finally we also note that (**A1**) subsumes (**A0**). Hence, we will not refer to (**A0**) separately.

Remark 2. **A2** can seem to be a strong assumption. However, in many applications, particularly those on social graphs, it is not difficult to satisfy this assumption. As a simple example, consider all devices which share a geographic location or IP, 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 link with fragmented identities based on confidence scores i.e. they have a probabilistic version of the adjacency matrix [87]. Such a method can usually be adapted to obtain a superset of neighbours with high probability (by including even low-confidence nodes as neighbours).

In addition to the assumptions (**A1**) and (**A2**), we will also

posit standard assumptions of network ignorability, positivity and consistency from causal literature [62]. All assumptions are formally stated in Appendix B.1. Since we are primarily considering a experimentation scenario, there do not exist any confounders. Moreover, positivity is easily ensured by choosing a good randomization scheme. Hence the assumptions (**A3**) to (**A5**) are naturally satisfied.

Standard Causal Assumptions	
Network Ignorability: $Y(\boldsymbol{z}) \perp\!\!\!\perp \boldsymbol{Z}\ \forall \boldsymbol{z}$	(**A3**)
Positivity: $P(\boldsymbol{z}\|\boldsymbol{X}) > 0\ \forall \boldsymbol{z}$	(**A4**)
Consistency: $Y_i = Y_i(\boldsymbol{z})$ if $\boldsymbol{Z} = \boldsymbol{z}$	(**A5**)

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 3. 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 $\mathcal{M}(i)$.

The joint distribution $p(\tilde{E}, E, Y|X, Z)$ factorizes as $p_\theta(Y|E, X)p(\tilde{E}|E)p(E|Z)$. We parameterize the outcome distribution $P(Y|E, X)$ via a GLM (Generalized Linear Model) which expresses the mean $\mathbb{E}[Y|Z = z, X = x]$ in terms of a neural network, i.e., we use a neural network for each of the functions c_0, c_1, g, w in (**A1**). For $p(\tilde{E}|E)$ we use a Gaussian model. Since the per-node allocations are independent, if $|\mathcal{M}(i)| \gg \mathcal{N}_i$, by law of large numbers this is a reasonable approximation for the error. Finally $p(Z|X)$ is just the allocation mechanism which is known to us as the experimenter (or can be estimated for observational data).

Technically, the latent variable E in this setting is a discrete variable, as Z is a binary assignment of treatments at individual devices. However, since the space is combinatorially large, we instead propose solving a continuous relaxation of the problem. We treat the latent variable as a continuous vector and use a variational method [42, 43] for estimation.

To use variational inference 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

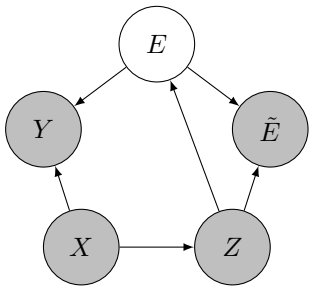


Figure 2: Graphical model depicting relationships between different variables for our model. Observed variables $\tilde{E}$ (noisy exposure), Y (effect/outcome), X (covariates), and Z (treatment allocation) are shaded to distinguish them from the hidden variable E (true treatment).

$\mathcal{L}_K$ [15], which is a lower bound for the conditional log-likelihood $p_\theta(x, y|z)$:

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

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 DReG estimator [93]. We also incorporated additional regularization on the neural network weights, and annealed the posterior q to match the prior $p(E|Z)$ for more stable training. Once the parameters θ have been learnt, τ can be estimated with the fitted outcome model $p_\theta(Y|E, X)$. More details on the estimation are included in Appendix B.4.

Remark 4. While the probability distribution can be arbitrarily parameterized with neural networks, all the networks used in our experiments are MLPs with one hidden layer and leaky ReLU activation. One can also use more powerful flow-based posteriors, instead of a Gaussian model.

Remark 5. Under GATE, we only consider $\boldsymbol{Z} = \vec{0}/\vec{1}$, and hence the latent exposure vector at test/prediction time is exactly known irrespectively of the exact graph.

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. We demonstrate the identifiability of our model, and state it as Proposition 4.1. The proof, included in the appendix, uses a result in Schennach and Hu [74]. Identifiability of the proposed probabilistic model was not previously known, and our result constitutes a new contribution to the field. We summarize the crux of the argument below, while deferring the details to Appendix B.

Proposition 4.1. *Under certain technical conditions* [1] *on the function g, the conditional mean function* $\mathbb{E}[Y|Z = z, X = x] = \mu_Y(x, z)$ *in our model is identifiable.*

When the graph $\boldsymbol{A}$ is exactly known, then under Assumptions (**A1**) to (**A5**) the problem of treatment effect estimation becomes a model fitting problem. Specifically, since the graph $\boldsymbol{A}$ is known, one can compute the exposures e_i, and then one can conduct a propensity weighted 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.

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 latent variable regression 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 (**A2**), $e_i(\mathcal{M})$ can be decomposed as $e_i(\mathcal{N}_i) + \Delta e_i$, where Δe_i is an independent error term. Thus, $e_i(\mathcal{M})$ act as noisy estimates of $e_i(\mathcal{N}_i)$.

Note that we have converted the problem of estimating the counterfactual functions to a problem of noisy regression with errors in covariates. While in general the broader family of noisy regression is unidentifiable, models of the proposed form:

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

can be shown to be identifiable from only the joint observations of $Y, \tilde{E}$ [74].

Remark 6. 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, our approach becomes equivalent to regression with endogenous covariate error, which requires additional information [106, 118].

5. Experiments

Before we describe the experiments and their results, we mention a few key research questions, and how our experiments are considered to answer each one of them.

RQ1 How does HIFIVE behave when all assumptions are satisfied?

RQ2 Is HIFIVE robust to violating Assumption (**A1**)?

RQ3 How sensitive are the results w.r.t Assumption (**A2**) ?

RQ4 Does HIFIVE work on real observational data?

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

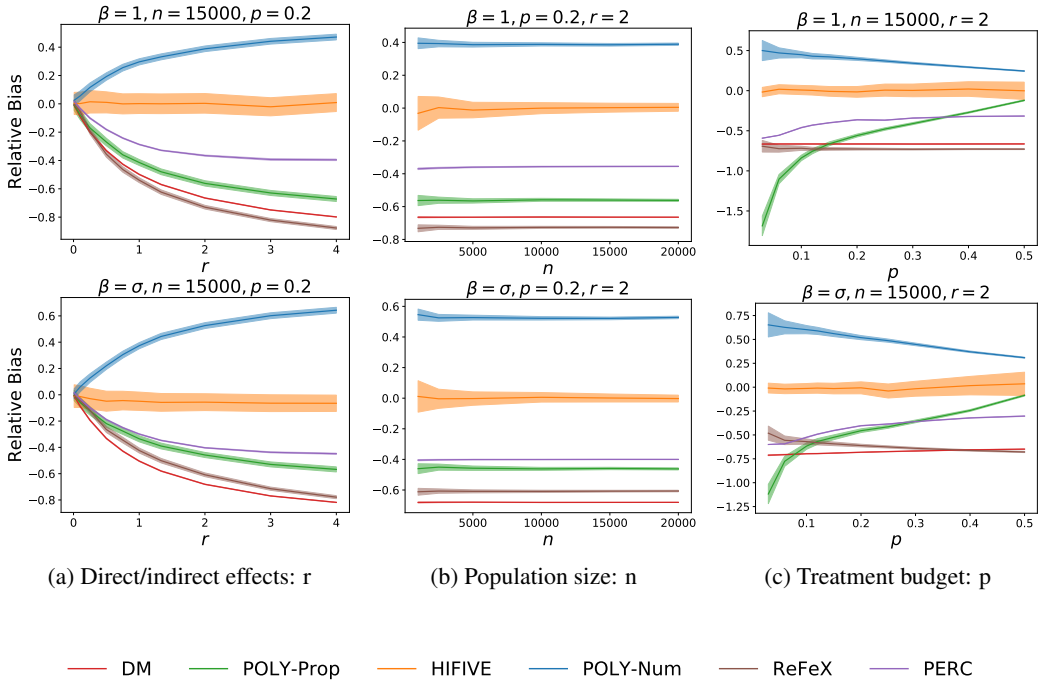
Experimental Section	Research Question
Section 5.1	Experimental Validity/RQ1
Section 5.2	Robustness/RQ2
Section 5.3	Uncertainity/RQ3
Section 5.4	Observational Data/RQ4

5.1. Synthetic Graphs

In this section, we first experimentally demonstrate the validity of HIFIVE by experimenting with synthetic data obtained from a model which satisfies our assumptions exactly. We experiment with Erdős-Rényi graphs to compare the performance of our estimator with other estimators. We simulate 100 different random graphs and run repeated experiments on each graph with random treatment assignments. Approximate neighbourhoods ($\mathcal{M}(i)$) are obtained by randomly adding nodes to node i's true neighbourhood $\mathcal{N}_i$. The covariates X are sampled from a multivariate normal distribution. Note that these are unit covariates and have no other connection to the graph or neighbourhoods. The potential outcomes $Y_i(\boldsymbol{z})$ are obtained by applying a function g on the exposure and adding a mean zero noise. The exposure are computed using the procedure in Cortez et al. [21]. We experiment with both a linear and non-linear (sigmoid-scaled linear) 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. Similar to Cortez et al. [21] we measure the strength of interference r as the ratio of norms of the self or direct influence and the indirect influence (more details in Appendix C.2).

We gauge the effectiveness of HIFIVE by benchmarking it against commonly employed estimators such as polynomial regression (Poly), ReFeX [36], PERC/DWR [117] and the difference-in-means (DM) estimators ($\hat{\tau}_{\text{SUTVA}}$). Except for the DM model, all other models need exact neighbourhoods, and so we use them in an oracle setting, i.e., they have access to the true graph. Due to incorporating large neighbourhoods, Horwitz-Thompson estimators failed to yield non-meaningful results in these trials.

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. Furthermore, even if the oracle graph is known, heterogeneity can cause bias in vanilla polynomial regression [59]. 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.



(a) Direct/indirect effects: r (b) Population size: n (c) Treatment budget: p

Figure 3: Plots visualizing the performance of various GATE estimators under Bernoulli design on Erdős-Rényi networks for both linear and quadratic sigmoidal 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.

5.2. AirBnB Model

Next, we conduct experiments with a framework designed for the AirBnB vacation rentals domain [50]. The original model is for rental listings and their bookings for a two-sided marketplace. We adapt this framework 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 one or more devices and, if interested, clicks on the ad but on only one device. This leads to interference between outcomes on the devices as only one (if any) receives a click. The simulation uses a type matching model where the devices and person have a latent type, and the probability of clicking is higher if the types match [2]. The treatment is considered to be a better algorithm which increases the relevance of the ad. This is considered as scaling the probability of click on the treated unit by the parameter α. The underlying outcome model in this scenario cannot be written as an exposure model. As such, this is a good testbed for testing robustness of our model, since, like in the real-world, exposure models are just approximations to the unknown and complex actual interference function. We use the protocol specified in Brennan et al. [14] .

For baselines, 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. [14], while the HT estimator is the one described in Section 3. We also work with the PERC/DWR [117] and ReFeX [36] estimators, which also need oracle graphs. The results are presented in Figure 4.

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 exposure using the model in Brennan et al. [14]. However even that model will be biased as the ground truth is not an exposure model. From the result it is also clear that HIFIVE works as well as the Oracle Exposure model. Furthermore, even on the MSE metric our model performs comparably to the Exp model. The results show that HIFIVE works even when the outcomes do not obey the assumed exposure mapping.

5.3. Effect of Network Uncertainty

Next we examine the impact of the neighbourhood accuracy $\mathcal{M}(i)$ in estimation. We experiment with Erdős-Rényi graphs as well as with the AirBnB Model. For these experiments, we fix a single graph, and compute the treatment

[2]Details in Appendix C.1

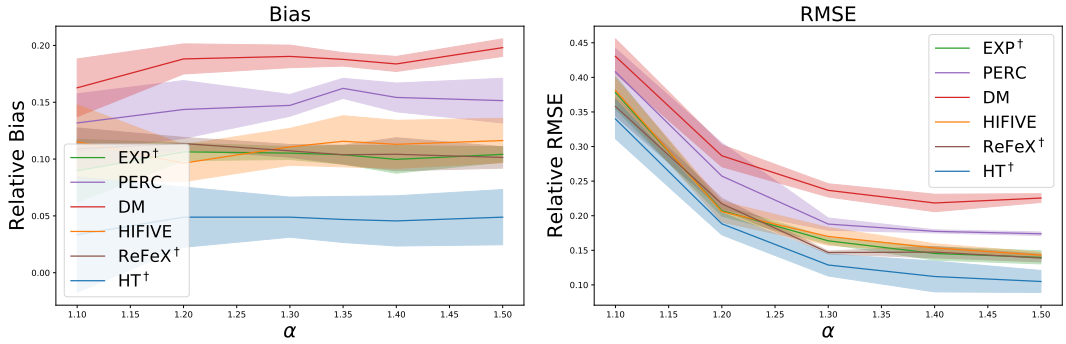
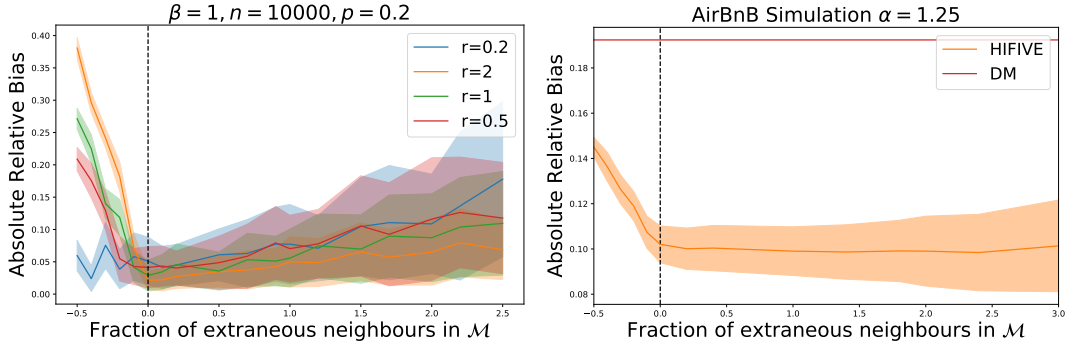


Figure 4: Visualization of performance of different GATE estimators on the AirBnB model. The lines represent a) absolute relative bias $|\frac{\hat{\tau}-\tau}{\tau}|$ and b) relative RMSE of various algorithms as the indirect treatment effect α increases. † indicates that the model has oracle access to true graph.



(a) Erdős-Rényi Networks

(b) AirBnB Simulator

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

effect estimate from HIFIVE 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, HIFIVE can offer an lower bias estimate of the treatment effect. Nonetheless, as the number of extraneous connections 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 neighbourhood sizes expand, bias diminishes while variance increases.

5.4. Application: Assessing Impact of Power Plant Emissions Controls

Next, we demonstrate an application of HIFIVE on observational data. We focus on estimate the effect of adoption of pollution reduction technologies at power-plants [59]. As ambient pollution is heavily influenced by spatially adjacent sources of pollution, adjusting for interference is important. We work with a public dataset on power generation facilities in USA used in Papadogeorgou et al. [59]. The outcomes Y_i correspond to measured pollutant levels, the devices correspond to the power plants, with treatment $Z = 1$ corresponding to adopting SCR/S-NCR technologies. To ensure comparability of the neighbourhoods with earlier methods, we used the clustering from [61] as the true neighbours, and for each unit added the two nearest non-true neighbours. The neighbourhood sizes range from 5 to 28, with median size of 10. We use the DM, EXP [4], Poly and ReFeX estimators as baselines, of which the latter three need true

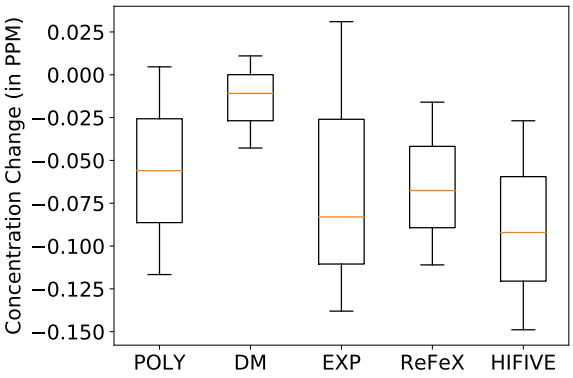


Figure 6: GATE estimates on ambient pollutant levels of adopting of SCR/SNCR technologies. The box plot depicts the mean and the 95% confidence interval. Note that all methods except ours use exact neighbourhoods.

neighbourhoods [3]. As this is observational data, we do not know the ground truth effect, and consider the EXP model as a reference. Estimation of propensity scores is done following [61]. Figure 6 shows that HIFIVE provides similar estimates as the the EXP and RefeX method, but does so without having exact neighbourhood information.

6. Conclusion

Identity fragmentation is an increasingly relevant problem in analysing human preferences from online A/B testing. In this work, we develop a method to estimate human preferences without the knowledge of the online identities that belong to the user. This is done under the practically far more feasible requirement of identifying supersets of the user's identities. We propose a variational model to estimate the counterfactual outcomes, and theoretically show the identifiability of our model under this superset assumptions. We also empirically establish the validity of our method and conduct an analysis of the effect of violations of the underlying assumptions. A limitation of our work is that the variance of the estimate grows with the size of the neighbourhoods, and so for practical applications one needs to balance the risk of higher variance against potential bias. Future research direction include incorporating temporal data and longitudinal studies. ZmJkN

Acknowledgements

We thank Yash Chandak and Saayan Mitra for the discussions, feedback, corrections, and other contributions to this work. We would also like to thank Daniel Sheldon for suggestions relevant to this work; and Marius Minea for help with the final editing and proofreading of the manuscript.

[3] Under an exposure model the EXP [4] estimator is a version of the HT estimator

Impact Statement

As this work is focused on estimating causal effects in online A/B testing within the context of privatized device identities and a cookie-less internet, there are important privacy considerations. Since this work uses approximate neighbourhoods $\mathcal{M}_i$ and the accuracy of these models is connected to device linking, there can be concerns aligned with GDPR and other privacy regulations. Specifically, $\mathcal{M}_i$ has to be obtained while ensuring regulatory compliance and protecting user data. Our work also has potential implications for controlled trials. Often, subjects in medical trials might not have all of their medical records. This scenario has potential similarities with device linking, where available but unlinked medical information can be considered as devices. As such, our method or its generalizations have potential connections with the responsible handling of sensitive information. We recognize the potential broader impact of our research while underscoring the importance of ethical considerations.

References

[1] Amemiya, T. (1983). Non-linear regression models. *Handbook of Econometrics*, 1:333–389.

[2] Aral, S. and Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1639.

[3] Aronow, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1):3–16.

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

[5] Arpino, B. and Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, 55(4):1770–1780.

[6] Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies*, 5(2):37–51.

[7] Auerbach, E. and Tabord-Meehan, M. (2021). The local approach to causal inference under network interference. *arXiv preprint arXiv:2105.03810*.

[8] Banerjee, A. V., Chassang, S., and Snowberg, E. (2017). Decision theoretic approaches to experiment design and external validity. In *Handbook of Economic Field Experiments*, volume 1, pages 141–174. Elsevier.

[9] Bargagli-Stoffi, F. J., Tortù, C., and Forastiere, L. (2020). Heterogeneous treatment and spillover effects

under clustered network interference. *arXiv preprint arXiv:2008.00707*.

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

[11] Beran, R. (1997). Diagnosing bootstrap success. *Annals of the Institute of Statistical Mathematics*, 49(1):1–24.

[12] Bhattacharya, R., Malinsky, D., and Shpitser, I. (2020). Causal inference under interference and network uncertainty. In Adams, R. P. and Gogate, V., editors, *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 1028–1038.

[13] Braun, J. and Griebel, M. (2009). On a constructive proof of Kolmogorov's superposition theorem. *Constructive Approximation*, 30:653–675.

[14] Brennan, J., Mirrokni, V., and Pouget-Abadie, J. (2022). Cluster randomized designs for one-sided bipartite experiments. *Advances in Neural Information Processing Systems*, 35:37962–37974.

[15] Burda, Y., Grosse, R. B., and Salakhutdinov, R. (2016). Importance weighted autoencoders. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

[16] Cai, J., De Janvry, A., and Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108.

[17] Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC Press.

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

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

[20] Coey, D. and Bailey, M. (2016). People and cookies: Imperfect treatment assignment in online experiments. In *Proceedings of the 25th International Conference on World Wide Web*, WWW 16.

[21] Cortez, M., Eichhorn, M., and Yu, C. L. (2022). Graph agnostic estimators with staggered rollout designs under network interference. *Advances in Neural Information Processing Systems*, 35:7437–7449.

[22] Cortez-Rodriguez, M., Eichhorn, M., and Yu, C. L. (2023). Exploiting neighborhood interference with low order interactions under unit randomized design. *Journal of Causal Inference*, 11(1).

[23] Cox, D. R. (1958). *Planning of experiments*. Wiley.

[24] Dorie, V., Harada, M., Carnegie, N. B., and Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in Medicine*, 35(20):3453–3470.

[25] Eckles, D., Karrer, B., and Ugander, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1).

[26] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26.

[27] Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9(2):139–158.

[28] Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

[29] European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council.

[30] Fang, X., Chen, A. W., and Young, D. S. (2023). Predictors with measurement error in mixtures of polynomial regressions. *Computational Statistics*, 38(1):373–401.

[31] Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, pages 733–760.

[32] Gui, H., Xu, Y., Bhasin, A., and Han, J. (2015). Network A/B testing: From sampling to estimation. In *Proceedings of the 24th International Conference on World Wide Web*, pages 399–409.

[33] Guo, W., Yin, M., Wang, Y., and Jordan, M. (2022). Partial identification with noisy covariates: A robust optimization approach.

[34] Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press.

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

[36] Han, K. and Ugander, J. (2023). Model-based regression adjustment with model-free covariates for network interference. *Journal of Causal Inference*, 11(1).

[37] Hernán, M. A. and Robins, J. M. (2021). *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC.

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

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

[40] Johari, R., Li, H., Liskovich, I., and Weintraub, G. Y. (2022). Experimental design in two-sided platforms: An analysis of bias. *Management Science*, 68(10):7069–7089.

[41] Kephart, J. O. and White, S. R. (1992). *Directed-graph epidemiological models of computer viruses*, pages 71–102. World Scientific.

[42] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

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

[44] Kuurkova, V. (1991). Kolmogorov's theorem is relevant. *Neural Computation*, 3(4):617–622.

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

[46] LeSage, J. and Pace, R. K. (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

[47] Leung, M. P. (2019). Causal inference under approximate neighborhood interference. *arXiv preprint arXiv:1911.07085*.

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

[49] Leung, M. P. (2022). Causal inference under approximate neighborhood interference. *Econometrica*, 90(1):267–293.

[50] Li, H., Zhao, G., Johari, R., and Weintraub, G. Y. (2022). Interference, bias, and variance in two-sided marketplace experimentation: Guidance for platforms. In *Proceedings of the ACM Web Conference 2022*, pages 182–192.

[51] Li, W., Sussman, D. L., and Kolaczyk, E. D. (2021). Causal inference under network interference with noise. *arXiv preprint arXiv:2105.04518*.

[52] Lin, T. and Misra, S. (2021). The identity fragmentation bias. *Available at SSRN 3507185*.

[53] Liu, L. and Hudgens, M. G. (2014). Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statistical Association*, 109(505):288–301. PMID: 24659836.

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

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

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

[57] Ogburn, E. L., Sofrygin, O., Diaz, I., and Van der Laan, M. J. (2017). Causal inference for social network data. *arXiv preprint arXiv:1705.08527*.

[58] Ogburn, E. L. and Vanderweele, T. J. (2013). Bias attenuation results for nondifferentially mismeasured ordinal and coarsened confounders. *Biometrika*, 100(1):241–248.

[59] Papadogeorgou, G., Choirat, C., and Zigler, C. M. (2019a). Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. *Biostatistics*, 20(2):256–272.

[60] Papadogeorgou, G., Imai, K., Lyall, J., and Li, F. (2020). Causal inference with spatio-temporal data: Estimating the effects of airstrikes on insurgent violence in Iraq. *arXiv preprint arXiv:2003.13555*.

[61] Papadogeorgou, G., Mealli, F., and Zigler, C. M. (2019b). Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3):778–787.

[62] Pearl, J. (2009). *Causality*. Cambridge University Press.

[63] Pearl, J. (2012). On measurement bias in causal inference. *arXiv preprint arXiv:1203.3504*.

[64] Pöllänen, A. and Marttinen, P. (2023). Identifiable causal inference with noisy treatment and no side information. *arXiv preprint arXiv:2306.10614*.

[65] Pouget-Abadie, J., Aydin, K., Schudy, W., Brodersen, K., and Mirrokni, V. (2019). Variance reduction in bipartite experiments through correlation clustering. *Advances in Neural Information Processing Systems*, 32.

[66] Qu, Z., Xiong, R., Liu, J., and Imbens, G. (2021). Efficient treatment effect estimation in observational studies under heterogeneous partial interference. *arXiv preprint arXiv:2107.12420*.

[67] Quin, F., Weyns, D., and Silva, C. C. (2023). A/B testing: a systematic review.

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

[69] Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

[70] Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4):377–401.

[71] Saha Roy, R., Sinha, R., Chhaya, N., and Saini, S. (2015). Probabilistic deduplication of anonymous web traffic. In *Proceedings of the 24th International Conference on World Wide Web*.

[72] Sävje, F., Aronow, P. M., and Hudgens, M. G. (2021). Average treatment effects in the presence of unknown interference. *The Annals of Statistics*, 49(2):673–701.

[73] Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics*, 8:341–377.

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

[75] Schmidt-Hieber, J. (2021). The kolmogorov–arnold representation theorem revisited. *Neural Networks*, 137:119–126.

[76] Seshadhri, C., Kolda, T. G., and Pinar, A. (2012). Community structure and scale-free collections of Erdős-Rényi graphs. *Physical Review E*, 85(5):056109.

[77] Shankar, S. (2022). Multimodal fusion via cortical network inspired losses. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

[78] Shankar, S. and Sarawagi, S. (2018). Labeled memory networks for online model adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.

[79] Shankar, S. and Sheldon, D. (2021). Sibling regression for generalized linear models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 781–795.

[80] Shankar, S., Sheldon, D., Sun, T., Pickering, J., and Dietterich, T. G. (2020). Three-quarter sibling regression for denoising observational data.

[81] Shankar, S., Sinha, R., Chandak, Y., Mitra, S., and Fiterau, M. (2024). A/B testing under Interference with Partial Network Information. In *International Conference on Artificial Intelligence and Statistics*, pages 19–27. PMLR.

[82] Shankar, S., Sinha, R., Mitra, S., Sinha, M., and Fiterau, M. (2023a). 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)*.

[83] Shankar, S., Sinha, R., Mitra, S., Swaminathan, V. V., Mahadevan, S., and Sinha, M. (2023b). 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.

[84] Shankar, S., Thompson, L., and Fiterau, M. (2022). Progressive fusion for multimodal integration.

[85] Shpitser, I., Wood-Doughty, Z., and Tchetgen, E. J. T. (2021). The proximal ID algorithm. *arXiv preprint arXiv:2108.06818*.

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

[87] Sinha, R., Saini, S., and Anadhavelu, N. (2014). Estimating the incremental effects of interactions for marketing attribution. In *2014 International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC2014)*, pages 1–6. IEEE.

[88] Sussman, D. L. and Airoldi, E. M. (2017). Elements of estimation theory for causal effects in the presence of network interference. *arXiv preprint arXiv:1702.03578*.

[89] Swaminathan, A. and Joachims, T. (2015). Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755.

[90] Tchetgen Tchetgen, E. J. and VanderWeele, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75. PMID: 21068053.

[91] Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. (2020). An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*.

[92] Toulis, P. and Kao, E. (2013). Estimation of causal peer influence effects. In *International Conference on Machine Learning*, pages 1489–1497.

[93] Tucker, G., Lawson, D., Gu, S., and Maddison, C. J. (2018). Doubly reparameterized gradient estimators for Monte Carlo objectives. *arXiv preprint arXiv:1810.04152*.

[94] Ugander, J., Karrer, B., Backstrom, L., and Kleinberg, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 329–337.

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

[96] Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press.

[97] VanderWeele, T. J., Tchetgen Tchetgen, E. J., and Halloran, M. E. (2014). Interference and sensitivity analysis. *Statist. Sci.*, 29(4):687–706.

[98] Veitch, V. and Zaveri, A. (2020). Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *arXiv preprint arXiv:2003.01747*.

[99] Viviano, D. (2020). Experimental design under network interference. *arXiv preprint arXiv:2003.08421*.

[100] Wakefield, J. (2004). Non-linear regression modelling and inference. In *Methods and Models in Statistics: In Honour of Professor John Nelder, FRS*, pages 119–153. World Scientific.

[101] Wang, Y., Chakrabarti, D., Wang, C., and Faloutsos, C. (2003). Epidemic spreading in real networks: An eigenvalue viewpoint. In *22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings.*, pages 25–34. IEEE.

[102] Wang, Y., Ouyang, H., Wang, C., Chen, J., Asamov, T., and Chang, Y. (2017). Efficient ordered combinatorial bandits for whole-page recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

[103] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

[104] Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

[105] Wong, J. C. (2020). Computational causal inference. *arXiv preprint arXiv:2007.10979*.

[106] Wooldridge, J. M. (2012). *Introductory Econometrics: A Modern Approach*. Cengage Learning.

[107] Wu, C., Lin, W., Zhang, X., Zhang, Y., Xie, W., and Wang, Y. (2024). PMC-LLaMA: Toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*.

[108] Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. (2018). Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*.

[109] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., and Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

[110] Yi, G. Y., Delaigle, A., and Gustafson, P. (2021). *Handbook of Measurement Error Models*. CRC Press.

[111] Yin, M., Shi, C., Wang, Y., and Blei, D. M. (2021). Conformal sensitivity analysis for individual treatment effects. *arXiv preprint arXiv:2112.03493*.

[112] Yu, C. L., Airoldi, E., Borgs, C., and Chayes, J. (2022). Estimating total treatment effect in randomized experiments with unknown network structure. *Proceedings of the National Academy of Sciences*, 119(44).

[113] Yuan, Y., Altenburger, K., and Kooti, F. (2022). Causal network motifs: identifying heterogeneous spillover effects in A/B tests. In *Proceedings of the Web Conference 2021*, pages 3359–3370.

[114] Zhang, C., Mohan, K., and Pearl, J. (2023). Causal inference under interference and model uncertainty. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, pages 371–385.

[115] Zhang, J. and Bareinboim, E. (2021). Bounding causal effects on continuous outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12207–12215.

[116] Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2017). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *arXiv preprint arXiv:1711.11286*.

[117] Zhao, Z., Kuang, K., Xiong, R., and Wu, F. e. a. (2022). Learning treatment effects under heterogeneous interference in networks. *arXiv preprint arXiv:2210.14080*.

[118] Zhu, Y., Gultchin, L., Gretton, A., Kusner, M. J., and Silva, R. (2022). Causal inference with treatment measurement error: a nonparametric instrumental variable approach. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, Proceedings of Machine Learning Research, pages 2414–2424.

A. Related Work

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 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 [4, 7, 51, 99]. Typically linearity with respect to neighbouring treatments is also assumed [25, 49, 114, 102] but is not neccessary [88]. A limitation of these approaches is that they require complete knowledge of the network structure. While our approach also relies on imposing an exposure-based structure to the form of interference, however *we work with an incomplete knowledge of the network*.

Treatment effect estimation with unknown network interference has also been studied beginning with the seminal work of Hudgens and Halloran [38]. 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 and Tabord-Meehan [7], Bhattacharya et al. [12], Liu and Hudgens [53], Tchetgen Tchetgen and VanderWeele [90], VanderWeele et al. [97] 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 [25, 32]. However, this still requires information about the clusters, and is not applicable if multiple clusters of the required type do not exist. On the other hand, *our method can handle general unstructured graphs*. Finally, there are methods, which under restrictive assumptions, use SUTVA based estimates for one-sided hypothesis tests for treatment effect under interference [19, 6, 45].

Estimation without any side information: Recently, some methods have been proposed based on multiple measurements which can address the issue of interference[83, 21, 112] 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 [82]. As such, we aim to develop a *method which can work with only a single trial and/or observational data from an existing test*.

B. Estimation and Identifiability

Proposition B.1. If the neighbourhood proposed by $\mathcal{M}$ i.e. $\mathcal{M}(i)$ always contains the true neighbourhood $\mathcal{N}_i$, and is sufficiently larger than $\mathcal{N}_i$, then under the exposure assumption we can treat ΔZ as approximately gaussian.

Proof. Under Equation (**A2**) we can rewrite the exposure under $\mathcal{M}$ as:

$$e_i(\mathcal{M}) = \sum_{j\in\mathcal{M}(i)} \phi(z_j, X_i) = \sum_{j\in\mathcal{M}(i)\cap\mathcal{N}_i} \phi(z_j, X_i) + \sum_{j\in\mathcal{M}(i)-\mathcal{N}_i} \phi(z_j, X_i)$$

Now, since allocation of device level treatments are independent, $Z_i \perp\!\!\!\perp Z_j$, as well as its independent of X_i, the individual exposure terms $\phi(Z_j, X_i) \perp\!\!\!\perp Z_i$ for any $i \in \mathcal{M}(i) - \mathcal{N}_i$. If $|\mathcal{M}(i)| >> |\mathcal{N}_i|\phi(z_j, X_i)$, then the central limit theorem implies that the sum is approximately $\sum_{j\in\mathcal{M}(i)-\mathcal{N}_i} \phi(z_j, X_i)$ as $N(\bar{\phi}, |\mathcal{M}(i) - \mathcal{N}_i|Var(\phi)) \approx N(\bar{\phi}, |\mathcal{M}(i)|Var(\phi))$ □

B.1. Assumptions

Assumptions

Model: $Y_i(\boldsymbol{z}, x_i) = \mathbb{E}[Y|Z = \boldsymbol{z}, X_i = x_i] + \epsilon$
$= c_0(x_i) + c_1(x_i)z_i + g(w_i^T \sum_{j\in\mathcal{N}_i} \phi(z_j, X_i)) + \epsilon$ **(A2)**

Neighbourhood Superset: $\mathcal{M}(i) \supseteq \mathcal{N}_i$ **(A3)**

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

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

Consistency: $Y_i = Y_i(\boldsymbol{z})$ if $\boldsymbol{Z} = \boldsymbol{z}$ **(A6)**

B.2. Identifiability

Proposition B.2. Our model is identifiable if 1) $\forall x, \mu_Y(x, z)$ is continuously differentiable everywhere as a function of z, and 2) $\forall x, \partial_z \mu_Y(x, z) \neq 0$

Before arguing the previous proposition, we first state Theorem 1 from [74]. Our presentation of this result broadly follows that of Pöllänen and Marttinen [64].

Theorem 1 from Schennach and Hu [74]: Let y, z, z^*, Δz, Δy be scalar real-valued random variables related through

$$y = g(z^*) + \Delta y \tag{3}$$
$$z = z^* + \Delta z, \tag{4}$$

and y, z are observed while all remaining variables are not and satisfy the following conditions:

Condition 1. The variables z^*, Δz, Δy, are mutually independent, $\mathbb{E}[\Delta z] = 0$, and $E[\Delta y] = 0$ (with $\mathbb{E}[|\Delta z|] < \infty$ and $\mathbb{E}[|\Delta y|] < \infty$).

Condition 2. $\mathbb{E}[e^{i\xi\Delta z}]$ and $\mathbb{E}[e^{i\gamma\Delta y}]$ do not vanish for any $\xi, \gamma \in \mathbb{R}$, where $i = \sqrt{-1}$.

Condition 3. (i) $\mathbb{E}[e^{i\xi z^*}] \neq 0$ for all ξ in a dense subset of $\mathbb{R}$ and (ii) $\mathbb{E}[e^{i\gamma g(z^*)}] \neq 0$ for all γ in a dense subset of $\mathbb{R}$ (which may be different than in (i)).

Condition 4. The distribution of z^* admits a uniformly bounded density $f_{z^*}(z^*)$ with respect to the Lebesgue measure that is supported on an interval (which may be infinite).

Condition 5. The regression function $g(z^*)$ is continuously differentiable over the interior of the support of z^*.

Condition 6. $g'(z^*) \neq 0$ almost everywhere, and $f_{z^*}(z^*)$ is continuous and nonvanishing

Theorem B.3. *Let Condition 1-6 hold. Then the following holds:*

1. *$g(z^*)$ is not of the form*

$$g(z^*) = a + b\ln(e^{cx^*} + d) \tag{5}$$

for some constants $a, b, c, d \in \mathbb{R}$. Then, $f_{z^*}(z^*)$ and $g(z^*)$ (over the support of $f_{z^*}(z^*)$) and the distributions of Δz and Δy are identified.

2. *If $g(z^*)$ is of the form* (5) *then, neither $f_{z^*}(z^*)$ nor $g(z^*)$ in Model 1 are identified iff z^* has a density of the form*

$$f_{z^*}(z^*) = A\exp(-Be^{Cx^*} + CDx^*)(e^{Cx^*} + E)^{-F}, \tag{6}$$

with $c \in \mathbb{R}$, $A, B, D, E, F \in \mathbb{R}^+$

Next, we argue how Theorem B.3 implies Proposition B.2. Consider the conditional versions of our, i.e. consider the restricted version where the covariates X have been fixed. It is clear from Proposition B.1 and Assumption **A2** that Equations (3) and (4) are satisfied for this model. Condition 1 of Theorem 1 also follows from Proposition B.1 and Assumption **A2**.

Condition 2,3 are technical conditions satisfied by most distributions (including Gaussian, Uniform and exponential family distributions). Condition 4 is satisfied because $\tilde{E}|E$ is approximately normal. Furthermore it will also hold for a variety of bounded continuous distributions. Condition 5,6 hold from the assumption on μ_Y stated in the proposition. With the conditions of Theorem B.3 satisfied, the conditional mean function $\mathbb{E}[Y|Z.X = x]$ are identified based on Theorem B.3 except for when $\mu_Y(x, z^*)$ might be of the form $a + b\ln(e^{cz^*} + d)$.

Since the conditional means $\mu_Y(Z, X = x)$ is identifiable for all x, the overall function $\mu_Y(Z, X)$ is also identified.

B.3. Relation to Schennach and Hu [74]

Schennach and Hu [74] proposed estimating the function g in Equation 3 through the following optimization.

$$\arg\max_{g} \max_{f_1, f_2, f_3} \ln \int f_1(y - g(z^*)) f_2(z - z^*) f_3(z^*) dz^* \tag{7}$$

where f_1, f_2, f_3 are restricted to be probability densities. This method is effectively maximizing the log-likelihood of the observed data under a latent variable framework. The latent variable, denoted as z^*, is integrated out within the objective which is a normalized density. Comparing this equation with our Equation 2, it becomes apparent that these methods are related. Specifically, the log-likelihood in Equation 2; can be obtained from Equation 7 by replacing z^* with e and z by $\tilde{e}$. The two key differences between our objective and that of Schennach and Hu [74] is a) that our likelihoods model conditioned on covariates X, and b) we can use specifics form for the densities f_2, f_3 and c) instead of directly maximizing likelihood we are maximizing the ELBO. The first difference is natural as we are fitting conditional models, unlike Schennach and Hu [74]. The choice of specific densities is also an issue in our scenario. As the experimenter, we already know the data generating density f_3 function, and by Proposition B.1, f_2 is well approximated by a Gaussian. This eliminates the need to learn these densities for our problem. Finally, instead of computing the objective integral via MCMC and optimization, we are instead learning using stochastic variational bayes method. Given ideal conditions, such as fully flexible posteriors and exact optimization, our proposed method converges towards the same solution as that obtained by the method of Schennach and Hu [74].

B.4. Estimation

Here we describe obtaining the estimate of treatment effect $\hat{\tau}$ from the model learnt in Section 4.2. We note that the variational posterior q_ϕ is providing us the estimate of the latent exposures E, while the model $p_\theta(Y|E, X)$ is learning the outcome models. Specifically, since $p_\theta(Y|E, X)$ is a GLM-style model parameterizing the mean $\mu_Y(e, x)$ one can directly obtain the counterfactual mean functions from it. These estimated means can be then plugged in Equation 1 to obtain the treatment effect $\hat{\tau}$.

Under **A1**, this computation is further simplified by noting that output of c_0 is independent of the treatment z. Furthermore, we can see from **A1** that the mean $\mathbb{E}[Y|E, X]$ is direct sum of the output of the networks c_0, c_1, w when provided the corresponding inputs. As such one can directly obtain the treatment effect using the following equation:

$$\hat{\tau} = \frac{1}{n}\sum_{i}^{n} \hat{\mu}_Y(\vec{1}, x_i) - \hat{\mu}_Y(\vec{0}, x_i)$$
$$= \frac{1}{n}\sum_{i}^{n} \Big[c_1(x_i) + g(w(x_i)^T e_i(\vec{1}, x_i))\Big]$$

Here c_1, g, w etc are neural networks whose parameter was estimated in learning p_θ.

B.5. Statistical Inference

In general analytical formulas for non-linear models are difficult and use some form of approximation using estimating equation or quasi-likelihood[1, 100]. An alternative is to use bootstrap approaches [26]. We describe a method for conducting inference in both these approaches here.

B.5.1. Parametric Bootstrap

Algorithm 1 Parametric Bootstrap

Input: $\mathcal{D} = \{\{X, Y, Z\}_{1:n}, A\}$, Bootstraps B, Estimator $\mathcal{A}$
$\hat{\theta}, \hat{\tau} \leftarrow \mathcal{A}(\mathcal{D})$
for b from 1 to B **do**
  $Z_1^*, \ldots, Z_n^* \sim P_{\hat{\theta}}(Z_i|X_i)$
  $\boldsymbol{Z} = \{Z_1^*, \ldots, Z_n^*\}$
  $Y_1^*, \ldots, Y_n^* \sim P_{\hat{\theta}}(Y|X_i, \boldsymbol{Z})$
  $\hat{\theta}^{*b}, \hat{\tau}^{*b} \leftarrow \mathcal{A}(\{\{X, Y^*, Z^*\}_{1:n}, A\})$
end for

return $\hat{\tau}, \left(\hat{\tau}^{*1}, \ldots, \hat{\tau}^{*B}\right)$

Parametric bootstrap [96, 11] is a model based variation of classical bootstrap [26, 27], wherein the distribution of an estimator $\mathcal{A}$ is obtained by repeatedly appliyng $\mathcal{A}$ to simulated datasets whose distribution mirrors that of the original data. In the parametric bootstrap, the simulated datasets are generated based on $P_{\hat{\theta}}$, representing the parametric distribution with the estimated parameter $\hat{\theta}$. We describe the algorithm in Algorithm 1, where $\mathcal{A}$ is our overall procedure which fits the variational model and return the model parameters and the estimated treatment effect. Bootstrap methods, generally make fewer assumptions compared to purely asymptotic approaches, provide practically tight bounds and works naturally with variational inference based methods [104]. In context of variational inference it is also related to posterior predictive checks [70, 31].

The general idea of the approach is to a) consider the estimated parameters $\hat{\tau}, \hat{\theta}$ as the ground truth, b) generate replicates from the generative distribution (in this case reassigning the treatments at nodes and sample outcomes from the new treatments), c) run the estimator $\mathcal{A}$ on the replicates to obtain replicate estimates $(\hat{\tau}^*)$, and d) then treat the pair $(\hat{\tau}^*, \hat{\tau})$ analogously to $(\hat{\tau}, \tau)$ to approximate the distribution of the latter. Mathematically, if $\hat{\xi}_\gamma$ is the $1-\gamma$ quantile of $\hat{\tau}^*$, then the intervals for τ can be obtained as $[\hat{\xi}_{1-\frac{\alpha}{2}},\ \hat{\xi}_{\frac{\alpha}{2}}]$ [26] for the chosen confidence level α.

B.5.2. Linearized Model

We propose to linearize the assumption (**A1**) model around the estimated parameters, and consider fitting the outcomes via a square loss [4], i.e. we fit

$$(Y_i - \mu_Y(\boldsymbol{Z}, X_i) - \sum_{j \in \mathcal{M}(i)} \partial_{Z_j}\mu_Y(\boldsymbol{Z}, X_i))^2 \psi(\tilde{E}, E, X)$$

where ψ includes the rest of the terms in the likelihood. The variance of the estimate is then determined by the (uncentered) covariance matrix for a linear regression problem [89, 66]. Specifically the posterior variance for the prediction Y_i is upper bounded by

$$\sigma_\epsilon^2 + [\sum_k c_{ik}]^2 (p(1-p))^{-1} (\sum_{j \in \mathcal{M}(i)} Z_j)^2$$

where c_{ik} are the coefficients of Z_k in the regression. We refer the readers to Theorem 3 in Qu et al. [66] for the derivation. In our case, the regression is derived from locally linearizing the $\mathbb{E}[Y|\boldsymbol{z}, X]$, and so the coefficient are nothing but the partial derivatives of the mean outcome function Y_i w.r.t Z_k. For the value of these derivatives, we can use the the current value of θ as the estimate. Next, for the variance of the effect τ, we see that the estimator is just the mean of n sample means of these $Y_i's$. If the max-degree of each node is bounded, then by generalized CLT [3, 47], the estimator is asymptotically normal with variance given by:

$$\frac{2\sigma_\epsilon^2}{n} + \frac{2}{n}\sum_i \left[[\sum_k c_{ik}]^2 (p(1-p))^{-1} (\sum_{j \in \mathcal{M}(i)} Z_j)^2 \right]$$

. If the max degree of any node in the graph is δ, then above sum can further be bounded by:

$$\frac{2\sigma_\epsilon^2}{n} + \frac{2}{n}\sum_i \left[[\sum_k c_{ik}]^2 (p(1-p))^{-1} \delta^2 \right]$$

This variance can then be used to provide conservative intervals for a Wald test [104]. Note however that this is only under a linearized approximation and hence using the above variance for confidence intervals are only approximately valid. However from the results of Sussman and Airoldi

[4] consider only a single unit i currently

[88], Cortez-Rodriguez et al. [22], this bound is minimax optimal in its dependence on p, σ, δ. As such these can still provide consistently conservative confidence intervals.

C. Experimental Details

C.1. AirBnB Model

The model used in these experiments is a version of the buyer and listing simulator developed by Li et al. [50]. The original model is a simulator for rental listings and their bookings for a two-sided marketplace scenario, with treatments affecting which seller listings are applied to by a buyer.

We adapt this simulator for our purposes, replacing customers with devices and listings with users. Each device and customer have a latent category, and the probability of watching an ad is significantly higher if the user and device category match. We assume that the user watches all ads that it has decided to watch, and then with a certain probability clicks on only of the ads it sees. Effectively, there is no temporal component in the treatments. The observed outcome (Y_i) is 1 if device i successfully receives a click on the ad. Since only one click is possible, the more ads are watched by the user, the lesser is the click rate at a single device, leading to interference. In our terminology, the experimental units are devices and the interference units are the users. The outcomes at an experimental unit is influenced by other experimental units that are incident on the the same interference unit. Consistent with prior literature [50, 40, 14], we use a 20 latent type matching model. To match consideration probabilities as mentioned in Brennan et al. [14] the ad-watching probability under the control assignment is 0.016 if the device and user share the same type. Similarly, the click probability was set to match acceptance probabilities mentioned in Brennan et al. [14]. The treatment tested is a recommendation algorithm, which increases the probability that a user watches an ad on the treated device.

C.2. Synthetic Graphs

The Erdos-Renyi (ER) model is commonly used for analyzing interaction networks in various experimental settings, particularly in the realm of social media [76] and epidemic control [41, 101]. In social media platforms, where connections form organically, ER graphs provide a reasonable simulation of how friendships, followerships, or interactions might evolve in an online community [28]. Additionally, in the context of epidemic control, ER graphs are valuable for studying disease spread [101].

We sample different random ER Graphs and run repeated experiments on these graphs with randomized bernoulli treatment assignment. For obtaining approximate neighbourhoods $\mathcal{M}(i)$ we compute another ER graph G' but clip the degree of the nodes to be between 10 and 100. Next we add the edges from G' to the original graph. The baselines include the POLY(Prop/Num) estimator is a polynomial regression on the exposure as computed by the fraction/number of treated nodes in the neighbourhood. The DM estimator signifies the classic difference in mean/ SUTVA estimator which is is simply the average outcomes on treated vs un-treated units. The ER graphs are made with an expected neighbourhood of size 20. The outcome model is similar to the potential outcomes model as in [21]:

$$Y_i(\boldsymbol{z}) = c_{i,\emptyset} + \sum_{j \in \mathcal{N}_i} \tilde{c}_{i,1} z_j + \sum_{\ell=2}^{\beta} \left(\frac{\sum_{j \in \mathcal{N}_i} \tilde{c}_{ij,2} z_j}{\sum_{j \in \mathcal{N}_i} \tilde{c}_{ij,2}} \right)^{\ell}, \quad (8)$$

where $i \neq j$, $\tilde{c}_{ij,2} = v_{i,2}|\mathcal{N}_i| / \sum_{k:(k,j)\in E} |\mathcal{N}_k|$. The coefficient $c_{i,\emptyset}, \tilde{c}_{i,1}, v_{i,2}$ are obtained as a linear function of the covariates X_i. For the non-polynomial model $\beta = \sigma$, we use the linear interference term $\frac{\sum_{j \in \mathcal{N}_i} \tilde{c}_{ij,2} z_j}{\sum_{j \in \mathcal{N}_i} \tilde{c}_{ij,2}}$ and scale it with a $x\sigma(x)$ function where σ is the sigmoid function.

C.3. Power Plant Emission Experiments

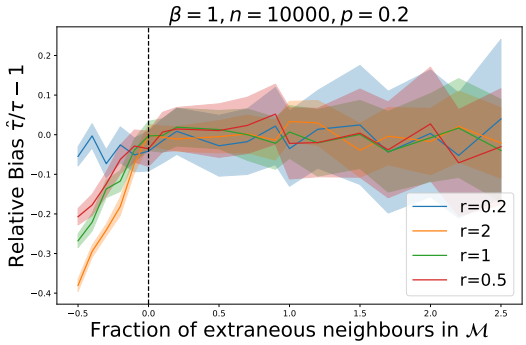
Selective Catalytic Reduction (SCR) and Selective Non-Catalytic Reduction (SNCR) are effective emission reduction technologies used in industrial settings, and their effectiveness in pollution has been supported in literature [59]. As ambient pollution is heavily influences by spatio-temporally adjacent sources of pollution, interference is a key component in the study of air pollution. We employ the identical dataset as Papadogeorgou et al. [59] to appraise the effect of SCR/SNCR adoption on ambient NOx, ozone and other gas levels. This openly accessible dataset encompasses 473 coal or gas-fired power generation facilities in USA. The dataset provides covariate details encompassing power plant characteristics, weather conditions, and demographic information in the surrounding regions. Due to the knowledge of geographical proximity, spatial-interference aware estimation methods can be used to provide plausible estimates of the treatment effect [60]. The POLY(Prop/Num) estimator is a polynomial regression on the exposure as computed by the fraction/number of treated nodes in the neighbourhood. The EXP estimator is the augmented inverse propensity estimator used by Papadogeorgou et al. [60], using a spatial exposure model. DM is the direct difference in mean estimate. For the DM estimate, a unit is considered treated if the nearest power-station adopted pollution reduction measures. Since this is an observational dataset, we do not know the propensities or the exact neighbourhood. To be consistent with earlier literature [59, 60], oracle neighbourhoods of the facilities are obtained according to Ward's method [103]. The covariates consists of 13 variables information on power plant characteristics, weather and demographic information of the locality. Since this is an observational model, we do

not have access to true probability of treatments. Following standard practice, we estimate this from the data. Specifically, the propensity scores were obtain via a logistic model with cluster specific random effects [59, 5].

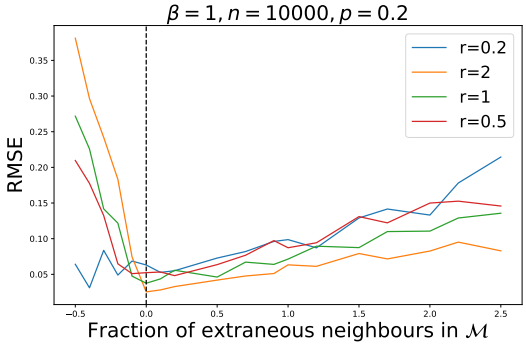
D. Additional Experimental Results

D.1. Effect of Network Uncertainty

In Figure 7 , we plot the relative bias and rmse of our estimator as the number of extraneous neighbours increase, for different levels of the strength of interference r.



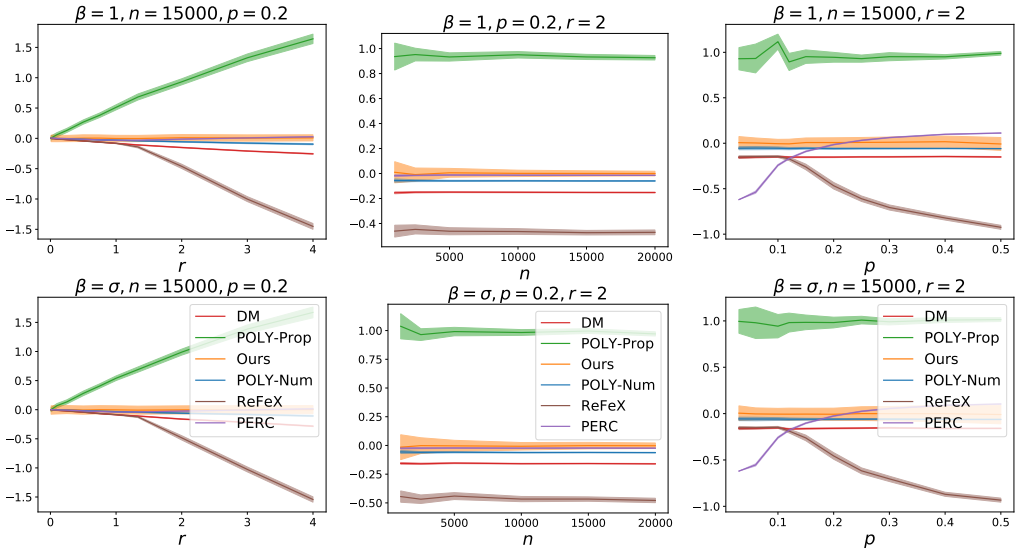
(a) Bias



(b) RMSE

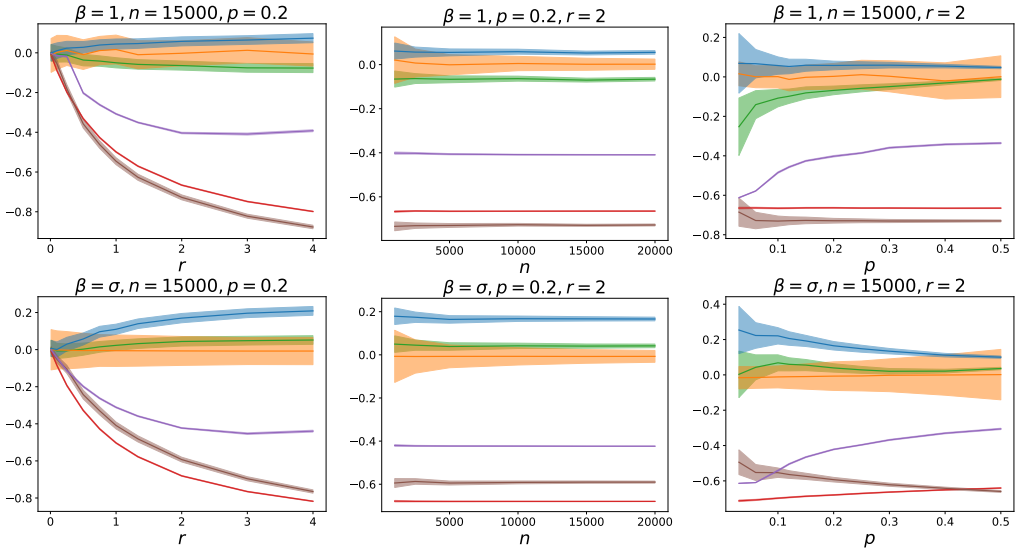
Figure 7: Visualization of the impact of neighbourhood sizes on GATE estimation. Negative fraction of neighbours indicate the case when $\mathcal{M}(i) \subset \mathcal{N}_i$ i.e. we missed pertinent neighbours. The bias tends to be high when gives small neighbourhoods, as they miss pertinent edges. As the neighbourhood sizes increase, the bias reduces, but the uncertainty widens.

D.2. Experiments on other random graphs



(a) Direct/indirect effects: r (b) Population size: n (c) Treatment budget: p

Figure 8: Plots visualizing the performance of various GATE estimators under Bernoulli design on Stochastic Block networks for both linear and sigmoidal 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.



(a) Direct/indirect effects: r (b) Population size: n (c) Treatment budget: p

Figure 9: Plots visualizing the performance of various GATE estimators under Bernoulli design on Strogatz-Watts networks for both linear and quadratic sigmoidal 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.