ONLINE EXPERIMENTATION UNDER PRIVACY INDUCED IDENTITY FRAGMENTATION

Shiv Shankar
University of Massachusetts
sshankar@cs.umass.edu

Ritwik Sinha
Adobe Research
risinha@adobe.com

Madalina Fiterau
University of Massachusetts
mfiterau@cs.umass.edu

ABSTRACT

Randomized online experimentation is a key cornerstone for evaluating decisions for online businesses. The methodology used for estimating policy effects in online experimentation is critically dependent on user identifiers. However, nowadays consumers routinely interact with online businesses across multiple devices which are recorded with different identifiers to maintain privacy. 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. Experiments show that estimators obtained through our procedure are 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. Users nowadays 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 [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). This means that a customer's effective persona as seen by the advertiser is broken into multiple units – a phenomenon known as 'identity fragmentation'[18, 50].

Identity fragmentation *across devices creates a fundamental issue in A/B testing, as the users' exposure to treatment becomes uncertain*. 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 a given device is the treatment group or the control group, and the user gets shown different ads on different visits, causing the observed outcome to be affected by mixed treatments.

Since the outcomes are dependent on user-level treatments, while our observation of them is at device level, we see treatments at a device affecting outcomes for other device. This has been known as interference. or spillover [36, 44]. Most methods involving spillover, assume strong restrictions on the structure of spillover [55, 46]. The deprecation of identifiers *introduces a new scenario, requiring the estimation of treatment effects on an uncertain network structure*. This is because while device linking might be difficult or impermissible, some information about the device graph can be obtained, for instance, from devices with signing information, or geolocation based on IP addresses or other

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

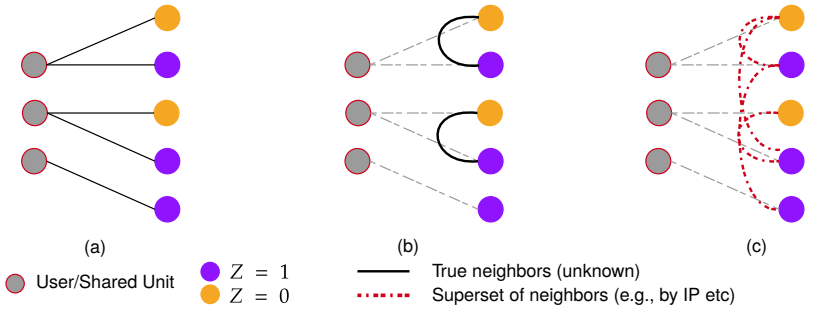


Figure 1: The bipartite graph (left) presents the connections between the set of users and devices. Treatments Z_i applied on a device, exposes the user of the device to the corresponding experience or algorithm etc. The outcomes depend on the total exposure a user has to the treatment, hence the outcome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

meta-data [66]. As such *an assumption can reasonably be made concerning the partial information about the device-user pairings, represented by the 'device graph'*.

Contribution We consider *global average treatment effect* (GATE) estimation under identity fragmentation *assuming 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 neighborhood' 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 show that the *GATE is identifiable in this setting*, and propose a variational inference based method to estimate the effect. 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 Existing works on network inference incorporate various sets of assumptions on the interference structure to provide an estimate of treatment effects [8, 14, 16, 30, 81]. 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 have been proposed based on multiple measurements which can address the issue of interference [73, 19, 98] without any further knowledge about its structure. However, such methods assume stationarity i.e. the outcomes do not vary between the trials, which 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 [72]. As such, 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, the only method which 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 detailed survey of the relevant literature is in the Appendix.

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 [54, 64], 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. Additionally we may have access to covariates X_i at the devices. Note that the devices

	General Graph	Uncertain Edges	Non-Linear Outcome	Single Trial
[36, 51]	×	✓	✓	✓
[99, 98]	✓	×	×	✓
[19, 73]	✓	✓	✓	×
[4, 67]	✓	×	✓	✓
[81, 23, 77]	✓	×	×	✓
Ours	✓	✓	✓	✓

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

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 [21, 36]; and is commonly called interference or spillover.

The user-device graph induces a dependence between device level outcomes. This dependence can be represented in a device-device 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 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). This means that the interference is limited to a node's neighbours in the device-device graph. 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 $\tau(\vec{1}, \vec{0}) = \frac{1}{n}\sum_{i=1}^{n} Y_i(\vec{1}) - \frac{1}{n}\sum_{i=1}^{n} Y_i(\vec{0})$. If the true graph $\boldsymbol{A}$ is known, under certain assumptions one can estimate the above treatment effect [36, 33]. 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 $\mathcal{M}$ 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 method is agnostic to how $\mathcal{M}$ was formed, and so in this work, we consider $\mathcal{M}$ as given. Often time, sume 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 neighborhood.

Our primary focus is on estimating the Generalized 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. A discussion of some common estimators like Horvitz-Thompson (HT) estimate and Difference-in-Means (DM) estimator, and their inapplicability is in the appendix.

4 METHOD

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 [36, 2, 4, 12]. Under this approach, one uses exposure variables e_i, and assumes that the outcome Y_i depends on the treatment $\boldsymbol{z}$ only via the exposure variable e_i i.e. $Y_i = Y_i(e_i(\boldsymbol{z}))$. Common examples include exposures measured as fraction[23, 81] or number [83] of neighbours receiving treatment. 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**).

Exposure Assumptions

Exposure Model: $Y_i(\boldsymbol{z}, x_i) = c_0(x_i) + c_1(x_i)z_i + g(w_i^T \sum_{j \in \mathcal{N}_i} \phi(z_j, X_i)) + \epsilon$ **(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 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. In addition to the Assumptions (**A1**) and (**A2**), we will also posit standard assumptions of network ignorability, positivity and consistency [60]. [2].

Remark 1. *Unlike most exposure models, we allow ϕ to be a vector function instead of scalar. Due to using vector ϕ,* (**A1**) *can support set function of neighbourhood treatments [11, 70, 42].*

Remark 2. ***A2*** *can seem to be a strong assumption, however in many applications, reasonable methods 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 [76, 66].*

4.1 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 $\tilde{E}$. Due to noisy treatment values, do-calculus rules [60] are not sufficient for identification [68].

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 the $p(\tilde{E}|E)$ we use a Gaussian model. Finally $p(Z|X)$ is just the allocation mechanism which is known to us as the experimenter.

Since the space of the latent variable E is combinatorially large, we solve a continuous relaxation of the problem, using variational inference [40, 41]. We use a Gaussian variational approximation with both mean and variance parameterized, as the posterior q_ϕ for the latent variable.

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 $\mathcal{L}_K$ [13], which is a lower bound for the conditional log-likelihood.

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

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 [82]. Once the parameters θ have been trained, τ can be estimated with the fitted outcome model $p_\theta(Y|E, X)$.

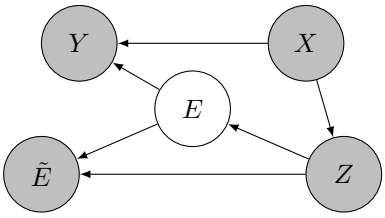


Figure 2: Graphical model depicting relationships between different variables for our model. Observed variables $\tilde{E}$ (noisy exposure), Y (effect/outcome), X (covariates), and Z (treatment allocation) are shaded to distinguish them from the hidden variable E (true treatment).

[2] discussed in Appendix

4.2 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. We demonstrate the identifiability of our model, and state it as Proposition 1. The proof, included in the appendix, uses a result in Schennach and Hu [69]. We summarize the crux of the argument below, while deferring the details to Appendix B.

Proposition 1. *Under certain technical conditions* [3] *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, one can compute the exposures e_i, and conduct a regression of the observed outcomes Y_i on the exposures e_i. Under standard assumptions [60] this identifies the population-level mean potential outcomes functions, denoted as μ_Y [16].

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 $e_i(\mathcal{N}_i)$.

Next, we argue the identifiability of the above regression task. In general models of the form:

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

are identified from observations of $Y, \tilde{E}$ [69]. Using a similar argument, our model is identifiable .

Remark 3. *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 is equivalent to regression with endogenous errors, which requires additional information [94, 104].*

5 EXPERIMENTS

5.1 AIRBNB MODEL

We conduct experiments with a model designed for the AirBnB vacation rentals domain [48]. The original model is for rental listings and their bookings for a two-sided marketplace. We adapt this model 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 its 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 treatment is considered to be a better algorithm which scales 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 in Brennan et al. [12] .

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. [12], while the HT estimator is the one described in Section 3. We also work with the PERC/DWR [103] and ReFeX [34] estimators, which also need oracle graphs. The results are presented in Figure 3.

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. [12]. However even that model will be biased as the ground truth is not an exposure model. 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. Our method works even when the outcomes do not obey the assumed exposure mapping.

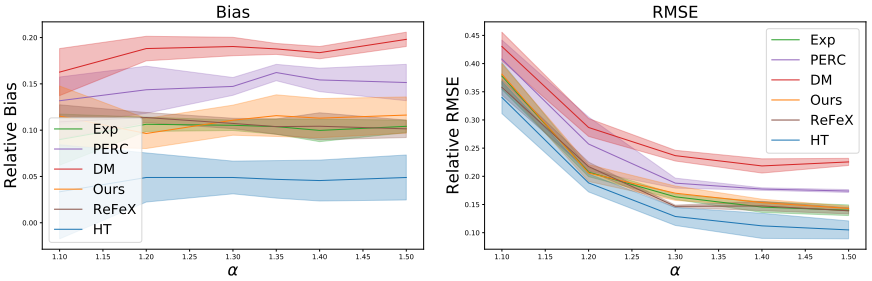
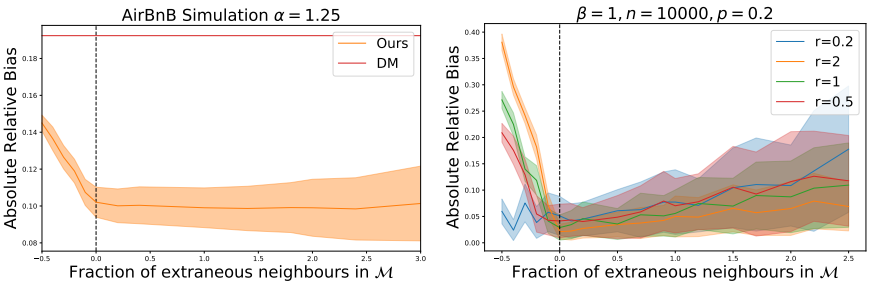


Figure 3: Visualization of performance of different GATE estimators on the AirBnB model. The lines represent a) absolute relative bias $|\frac{\hat{\tau}-\tau}{\tau}|$ and b) relative RMSE of various algorithms as the indirect treatment effect α increases.



(a) AirBnB Simulator

(b) Erdos-Renyi Networks

Figure 4: 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 is high when given small neighbourhoods, as they miss pertinent edges. As the $|\mathcal{M}(i)|$ increase, the bias reduces, but the uncertainty widens.

5.2 Effect of Network Uncertainty

Next we examine the impact of the neighborhood accuracy $\mathcal{M}(i)$ on our methos. We experiment with both the AirBnB Model and synthetic Erdos-Renyi graphs. For these experiments, we fix a single graph, and find how treatment effect estimate from our method behaves as we change the neighbourhoods $\mathcal{M}(i)$.

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 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 neighborhood sizes expand, bias diminishes while variance increases.

6 Conclusion

Identity fragmentation is an increasingly relevant problem in online A/B testing. We develop 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. We establish the efficacy of our method under this superset assumption. 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.

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

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] 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.

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

[10] 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. PMLR.

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

[12] 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.

[13] 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*.

[14] 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.

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

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

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

[18] 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.

[19] 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*.

[20] Cortez-Rodriguez, M., Eichhorn, M., and Yu, C. L. (2022). Exploiting neighborhood interference with low order interactions under unit randomized design. *arXiv preprint arXiv:2208.05553*.

[21] Cox, D. R. (1958). Planning of experiments.

[22] 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.

[23] 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).

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

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

[26] Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

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

[28] 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.

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

[30] 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. International World Wide Web Conferences Steering Committee.

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

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

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

[34] Han and Ugander (2023). Model-based regression adjustment with model-free covariates for network interference.

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

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

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

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

[39] Kephart, J. O. and White, S. R. (1992). Directed-graph epidemiological models of computer viruses. In *Computation: the micro and the macro view*, pages 71–102. World Scientific.

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

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

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

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

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

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

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

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

[48] 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.

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

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

[51] 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.

[52] 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.

[53] 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.

[54] 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).

[55] 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*.

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

[57] 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.

[58] 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*.

[59] 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.

[60] Pearl, J. (2009). *Causality*. Cambridge university press.

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

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

[63] 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*.

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

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

[66] 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*, WWW 15.

[67] 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.

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

[69] 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.

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

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

[72] 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)*.

[73] 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.

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

[75] 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.

[76] 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.

[77] 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*.

[78] Swaminathan, A. and Joachims, T. (2015). Batch learning from logged data through counterfactual risk minimization. *The Journal of Machine Learning Research*.

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

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

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

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

[83] 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. ACM.

[84] 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.

[85] Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

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

[87] 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*.

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

[89] 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.

[90] 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.

[91] 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.

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

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

[94] Wooldridge, J. M. (2009). *Econometrics: Modern Approach*. Cengage AU.

[95] 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*.

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

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

[98] 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*.

[99] 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.

[100] 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*. PMLR.

[101] Zhang, J. and Bareinboim, E. (2021). Bounding causal effects on continuous outcomes.

[102] 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*.

[103] Zhao, Z., Kuang, K., Xiong, R., and Wu, F. e. a. (2022). Learning treatment effects under heterogeneous interference in networks. *Neurips*.

[104] 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. PMLR.

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, 49, 88]. Typically linearity with respect to neighbouring treatments is also assumed [23, 47, 100, 91] but is not neccessary [77]. 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 [36]. 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. [10], Liu and Hudgens [51], Tchetgen and VanderWeele [79], VanderWeele et al. [86] 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 [23, 30]. 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 [17, 6, 43].

Estimation without any side information: Recently, some methods have been proposed based on multiple measurements which can address the issue of interference[73, 19, 98] 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 [72]. As such, we aim to develop a *method which can work with only a single trial and/or observational data from an existing test*.

Estimation with Noisy Data Many methods and heuristics have been proposed for estimation of treatment effect [15, 68, 56, 52] with measurement noise in data. Yi et al. [96] 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 [35], outcome [75], confounders [61, 53] and mediators [84]. Methods based on assuming knowledge of the error model are also common [32, 74, 28]. Existing methods for estimating causal effects under noise rely upon additional information such as repeated measurements [73, 19], instrumental variables [104, 80] or a gold standard sample of measurements [72]. While few works have also tried to study causal inference with measurement errors and no side information [53, 62], 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 [102, 95, 101, 97, 31] and sensitivity analysis [37, 87, 22].

Inverse Propensity/Horvitz-Thompson Estimate If the graph is known and when all treatment decisions independently set with probability p, one can use the classic Horvitz Thompson estimator (or inverse propensity estimator) as:

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

This inverse propensity estimators (and its derivatives) do not require any further assumption other than randomization and positivity. However, this estimator filters out any units for which all neighbours are not in control or treatment groups, and is not be meaningful, when there do not not exist

units for which all the neighbours are in control or treatment groups. This is particularly troublesome under lack of indentifiers, as uncertainty in the linkages 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

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

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.

B Estimation and Identifiability

Proposition 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 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 [69]. Our presentation of this result broadly follows that of Pöllänen and Marttinen [62].

Theorem 1 from Schennach and Hu [69]: Let y, z, z^*, Δz, Δy be scalar real-valued random variables related through

$$y = g(z^*) + \Delta y \tag{2}$$
$$z = z^* + \Delta z, \tag{3}$$

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

1. *$g(z^*)$ is not of the form*
$$g(z^*) = a + b\ln(e^{cx^*} + d) \tag{4}$$
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* (4) *then, neither $f_{z^*}(z^*)$ nor $g(z^*)$ in Model 1 are identified iff z^* has a density of the form*
$$f_{z^*}(z^*) = A\exp(-Be^{Cx^*} + CDx^*)(e^{Cx^*} + E)^{-F}, \tag{5}$$
with $c \in \mathbb{R}$, $A, B, D, E, F \in \mathbb{R}^+$

Next, we argue how Theorem 1 implies Proposition 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 1 and Assumption **A2** that Equations (2) and (3) are satisfied for this model. Condition 1 of Theorem 1 also follows from Proposition 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 1 satisfied, the conditional mean function $\mathbb{E}[Y|Z.X = x]$ are identified based on Theorem 1 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 [69] method

Schennach and Hu [69] proposed estimating the function g in Equation 2 through the following optimization.

$$g = \arg\max_{g} \max_{f1,f2,f3} \ln \int f1(y - g(z^*))f2(z - z^*)f3(z^*)dz* \tag{6}$$

where $f1, f2, f3$ 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 1, it becomes apparent that these methods are related. Specifically, the log-likelihood in Equation 1; can be obtained from Equation 6 by replacing z^* with e and z by $\tilde{e}$. The two key differences between our objective and that of Schennach and Hu [69] is a) that our likelihoods model conditioned on covariates X, and b) we can use specifics form for the densities $f2, f3$ 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 [69]. The choice of specific densities is also an issue in our scenario. As the experimenter, we already know the data generating density $f3$ function, and by Proposition 1, $f2$ 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 [69].

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 **??** 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:

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

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, 89].Another approach is to use bootstrap approaches [24]. 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 [85, 9] is a model based variation of classical bootstrap [24, 25, 26], 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 [93]. In context of variational inference it is also related to posterior predictive checks [65, 29].

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 re-assigning 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 [78, 63]. 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. [63] 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, 45], 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 [93]. 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 [77], Cortez-Rodriguez et al. [20], this bound is minimax optimal in its dependence on p, σ, δ. As such these can still provide consistently conservative confidence intervals.

[4] consider only a single unit i currently