Bayesian Estimation of Causal Effects Using Proxies of a Latent Interference Network

Bar Weinstein
Daniel Nevo {BARWEIN, DANIELNEVO}@TAUEX.TAU.AC.IL
Department of Statistics and Operations Research
Tel Aviv University

Abstract

Network interference occurs when treatments assigned to some units affect the outcomes of others. Traditional approaches often assume that the observed network correctly specifies the interference structure. However, in practice, researchers frequently only have access to proxy measurements of the interference network due to limitations in data collection or potential mismatches between measured networks and actual interference pathways. In this paper, we introduce a framework for estimating causal effects when only proxy networks are available. Our approach leverages a structural causal model that accommodates diverse proxy types, including noisy measurements, multiple data sources, and multilayer networks, and defines causal effects as interventions on population-level treatments. The latent nature of the true interference network poses significant challenges. To overcome them, we develop a Bayesian inference framework. We propose a Block Gibbs sampler with Locally Informed Proposals to update the latent network, thereby efficiently exploring the high-dimensional posterior space composed of both discrete and continuous parameters. The latent network updates are driven by information from the proxy networks, treatments, and outcomes. We illustrate the performance of our method through numerical experiments, demonstrating its accuracy in recovering causal effects even when only proxies of the interference network are available.

Keywords: Spillovers, Peer effects, Noisy networks, Network regression, Locally Informed Proposals.

1 Introduction

In causal inference it is commonly assumed that there is no interference between units, that is, the treatment assigned to one unit does not affect the outcomes of other units (Cox, 1958). However, researchers are often interested in settings where the treatment of interest can spill over from treated units to untreated ones, leading to interference between units. The term interference encompasses various forms of transmission mechanisms between interacting units, including social influence, information diffusion, contagion, and protection against infectious diseases. Researchers often seek to estimate the effects of a postulated treatment policy, e.g., vaccination allocation in a population. To estimate the influence of such policies under interference, researchers typically rely on detailed information about the network structure that encodes the structure of pairwise unit interactions.

However, the network that researchers use to encode the interference structure might not be the actual interference network for at least two reasons. First, accurately measuring

networks of social interactions is a formidable task. For example, social networks are often measured through self-reported elicitation of social relations, where respondents are asked to list the names of others with whom they have relevant social interactions. The challenge of correctly measuring social interactions is well known in the network science literature. Common remedies for this challenge include improvements in research design (Marsden, 1990), or the use of models that account for measurement error (Butts, 2003; Young et al., 2021; De Bacco et al., 2023).

Second, even when a network is measured without error, the observed social ties may not correspond to the specific pathways through which interference operates. Researchers often rely on networks that are available or convenient to measure, yet the actual interference might propagate through a different, unobserved structure. Namely, there might be a mechanism mismatch between the true network and the observed proxies rather than measurement reliability. For example, researchers may exploit available online social media connections, while in reality, physical contacts, which are more challenging to measure, are more relevant to that specific study. That is, the online connections (observed proxies) can be measured without noise, while the intervention (e.g., information about a new technology) might spread only through physical connections (true latent network). While the online networks are "correct" in that they accurately record online interactions, they are an imperfect representation of the actual interference structure (physical connections), and thus serve as proxies for it. They are distinct graphs, but are likely correlated due to shared latent social traits.

Thus, researchers often observe only proxy measurements of the actual interference network, which remains latent. Treating the observed proxy as the true interference structure can result in misspecification of the interference structure that can bias causal effect estimates (Weinstein and Nevo, 2026). Figure 1 illustrates such network misspecification in a small population. Despite this, it is common in the literature to assume that the measured network accurately represents the interference structure (Ugander et al., 2013; Aronow and Samii, 2017; Sofrygin and van der Laan, 2017; Forastiere et al., 2020; Tchetgen Tchetgen et al., 2020; Leung, 2022; Forastiere et al., 2022; Ogburn et al., 2024; Hayes et al., 2025).

In this paper, we propose a structural causal models (SCMs) framework for estimating causal effects under network interference, when only proxy measurements of the true interference structure are available. Such scenarios include the case of a single erroneously measured network, repeated measures of noisy networks, multiple networks from varying sources, and multilayer networks (see Section 3.1 for extended examples). The SCMs view the true interference network as a realization from a random network model. Crucially, our framework accommodates various structural configurations, including both causal and non-causal proxy mechanisms, and explicitly addresses scenarios where treatment assignment depends on either the latent interference network or the observed proxies. We focus on the causal effects of interventions on the population treatments, addressing various policy-related estimands.

We discuss the identifiability of the structural parameters in simplified settings, illustrating how the outcomes, treatments, and proxy networks jointly contribute to identification. Identifying these structural parameters allows for the recovery of the latent network's conditional distribution, motivating our proposed Bayesian framework.

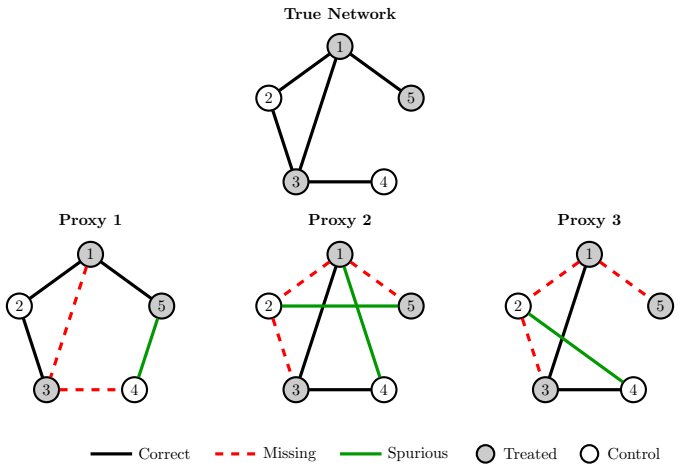


Figure 1: Network misspecification in a small population with binary treatments. "True Network" is the true latent network, while "Proxy 1-3" are three proxy networks with some missing and spurious edges. Each of the proxies provides partial and incomplete information about the true network. Nodes are colored by the binary treatment value of each unit. For example, node 1 is directly connected to two treated nodes (3 and 5) in the true network, while in each of the proxies, node 1 is connected only to one treated neighbor.

We develop a Bayesian inferential framework for estimation. Since the true interference network is latent, the posterior is composed of a mixed space of discrete and continuous components, where the discrete space (latent network) is large. That makes sampling from the posterior challenging and computationally expensive. To address this, we propose a Block Gibbs sampler that alternates between updating the continuous parameters and the discrete latent network. For the network updates, we employ Locally Informed Proposals (Zanella, 2020) to efficiently explore the high-dimensional discrete space, utilizing gradient-based approximations to render these proposals computationally feasible. This inference framework allows researchers to flexibly estimate the parameters of the various SCMs described above. We demonstrate the performance of our proposed methods through numerical illustrations on fully- and semi-synthetic data. Code is available at `https://github.com/barwein/BayesProxyNets`.

The remainder of this paper is organized as follows. In Section 2, we introduce our formal setup, present the SCMs, and define the causal estimands of interest. Section 3 categorizes the different types of proxy measurements and provides concrete examples of true and proxy network models. Section 4 investigates the identification of structural parameters in simplified settings. Section 5 describes the proposed Bayesian algorithm for posterior sampling and the causal effects estimation scheme. Section 6 demonstrates the performance of our method through numerical experiments. Finally, Section 7 concludes with a discussion of extensions and directions for future work.

1.1 Related Literature

A significant portion of the interference literature operates under the assumption that the interference network is fixed, correctly measured, and exogenous to the structural models (Sofrygin and van der Laan, 2017; Tchetgen Tchetgen et al., 2020; Ogburn et al., 2024). In contrast, our framework treats the interference network as a realization from a random network model, integrating it directly into the SCM. Other researchers have taken similar perspectives while assuming that the observed network accurately represents the interference structure (Goldsmith-Pinkham and Imbens, 2013; Li and Wager, 2022; Clark and Handcock, 2024).

When the assumption of correctly specified observed network is relaxed, existing solutions often rely on specific structural assumptions. A prominent line of work focuses on the linear-in-means model (Manski, 1993), developing corrections for specific error mechanisms such as random measurement error (Hardy et al., 2019; Boucher and Houndetoungan, 2022) or edge censoring (Griffith, 2021). These approaches typically leverage the parametric structure of the linear framework for identification. Other strategies require specific data availability, such as repeated noisy measurements to achieve identification via design-based estimators (Li et al., 2021), or rely on aggregating relational summaries when the full network is unobserved (Reeves et al., 2024). Alternatively, Toulis and Kao (2013) proposed modeling unobservable interactions through sufficient statistics of treated neighbors.

Beyond specific error corrections, our work relates to the broad literature on latent network inference. While probabilistic models for recovering latent networks from observed proxy networks are well-studied in the network science literature (Butts, 2003; Young et al., 2021; De Bacco et al., 2023), they are typically applied as distinct tasks, separate from causal effect estimation. A natural heuristic involves a two-stage approach: first inferring the network and then treating it as fixed for causal analysis. We show in the numerical experiments (Section 6) that such two-stage approaches can lead to estimation errors and underestimation of uncertainty. Our Bayesian framework improves upon this by jointly estimating the latent network and causal parameters, thereby properly propagating uncertainty from the network inference stage to the causal estimates.

Beyond causal inference under network interference, there is an emerging literature on regression problems with network-linked covariates. While this literature typically assumes the network is observed without error (Zhu et al., 2017; Li et al., 2019), exceptions exist that account for sampling uncertainty of the network (Le and Li, 2022). Although our primary focus is on causal estimation, our framework is adaptable to such network regression problems where only proxies are observed.

In network interference scenarios, treatments of other units are typically assumed to affect the potential outcome of one unit only through values of a summarizing exposure mapping (Ugander et al., 2013; Manski, 2013; Aronow and Samii, 2017). However, recent research has highlighted the challenges associated with this approach. Sävje (2024) demonstrated that incorrect specification of these mappings can introduce bias and obscure the interpretation of results. Building on this, Weinstein and Nevo (2026) showed that such biases may result from misspecification of the network structure itself. In this paper, we address these challenges by strictly distinguishing between modeling assumptions and causal targets. While we utilize exposure mappings as summarization tools to render the high-dimensional

outcome model tractable, we define causal estimands as interventions on population-level treatments (policies). This separation allows us to define and estimate meaningful static, dynamic, and stochastic policy effects even when precise unit-level exposures are uncertain, thus mitigating the interpretability issues raised by Sävje (2024). Other methodological advancements have focused on robust estimation using multiple proxy networks (Weinstein and Nevo, 2026), testing the exposure mapping specification through randomization tests (Athey et al., 2018; Puelz et al., 2022), or learning the exposure mapping from the data while assuming the observed network is correct (Ohnishi et al., 2022).

Finally, we note that certain causal effects can be estimated without network data. Sävje et al. (2021) showed that the Expected Average Treatment Effect (EATE), which is an effect marginalized over other units' treatments, can be consistently estimated with the common design-based estimators under limiting interference dependence between units. Yu et al. (2022) showed that the Total Treatment Effect (TTE) – treating all units versus none – can be unbiasedly estimated, under restrictions on the potential outcomes and the experimental design. In addition, Shirani and Bayati (2024) developed methods for estimating TTE using only an assumed distribution of the network. However, analyzing more granular policy interventions, such as the dynamic or stochastic policies considered in this paper, requires correct network measurements that are often unavailable to researchers, as discussed in Section 3.

1.2 Our Contribution

Our work advances the literature on causal inference with network interference through three primary contributions.

Flexible Structural Causal Models. We propose a general SCMs framework where the interference network is treated as a latent variable. This framework generalizes existing approaches by accommodating diverse structural configurations, including both *causal proxies* (e.g., noisy measurements) and *non-causal proxies* (e.g., multilayer networks sharing latent coordinates). Furthermore, it explicitly models distinct treatment assignment mechanisms, where treatment allocation depends either on the latent interference network or on the observed proxies.

Joint Estimation. We develop a unified Bayesian framework that jointly estimates the causal effects and reconstructs the latent interference network. This contrasts with heuristic two-stage approaches that first infer the network and then treat it as fixed, which often leads to underestimation of uncertainty. Our end-to-end estimation leverages information from all available modules (e.g., proxies, treatments, and outcomes) to recover the latent structure, while ensuring that uncertainty in the network is propagated to the final causal estimates.

Inference for Discrete Latent Networks. The joint estimation poses a significant computational challenge due to the high-dimensional discrete space of the latent network. We address this by developing a Block Gibbs sampler equipped with Locally Informed Proposals (LIP) (Zanella, 2020). We utilize gradient-based approximations to explore the posterior space, rendering Bayesian inference feasible for network sizes where brute-force marginalization is impossible. Furthermore, we provide a formal analysis of the approxima-

tion error, deriving a refined efficiency bound that exploits the structure of network updates to improve upon general global bounds (e.g., Grathwohl et al., 2021).

2 Settings and Causal Models

2.1 Setup and Notations

Consider a finite population of N units indexed by $i \in [N] = \{1, \ldots, N\}$. Denote the treatment vector of the entire population by $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$. Denote the observed outcomes by $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$. The interference network is represented by its $N \times N$ adjacency matrix $\boldsymbol{A}^*$. Denote by $\boldsymbol{A}_i^*$ the i-th row of $\boldsymbol{A}^*$. For simplicity, $\boldsymbol{A}^*$ is assumed to be unweighted and undirected, that is, $A_{ij}^* = A_{ji}^* = 1$ only if there is an edge between units i and j, and by convention $A_{ii}^* = 0$.

The true network $\boldsymbol{A}^*$ is not observed. Instead, we assume that researchers observe $B \geq 1$ proxy measurement(s) $\mathcal{A} = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_B\}$ of $\boldsymbol{A}^*$, such that each $\boldsymbol{A}_b$, $b \in [B] = \{1, \ldots, B\}$, is an adjacency matrix of the same dimension as $\boldsymbol{A}^*$. Section 3.1 provides detailed examples of possible proxies. Crucially, treatment interference is assumed to transmit through $\boldsymbol{A}^*$ and not through any of the proxies in $\mathcal{A}$. This assumption is formalized in the SCMs given below. Let $\boldsymbol{X} \in \mathbb{R}^{N \times q}$ be a matrix of baseline covariates, where $\boldsymbol{X}_i$ is the q covariates of unit i. The observed data are $\boldsymbol{\mathcal{D}} = (\boldsymbol{Y}, \boldsymbol{Z}, \mathcal{A}, \boldsymbol{X})$. Lastly, throughout the paper, we use $p(\cdot)$ and $p(\cdot \mid \cdot)$ as generic notations for (conditional) probability distributions.

2.2 Structural Causal Models

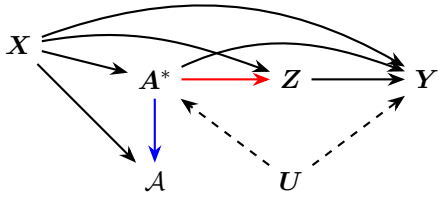
We use an SCM framework (Pearl, 2009), which models the data-generation mechanism through sequential evaluation of structural equations. Our proposed SCMs generalize previous models (Sofrygin and van der Laan, 2017; Ogburn et al., 2024) by including the interference network generation, as well as the proxy formation, as part of the model. The directed acyclic graphs (DAGs) in Ogburn et al. (2024) are essentially conditioned on a fixed network, while our SCMs treat the interference network as a realization from a random network model. That is, each DAG in Ogburn et al. (2024) can be viewed as a slice of our SCMs, given a specific network $\boldsymbol{A}^*$ realization. The underlying assumed structures can be represented by population-level DAGs, as illustrated in Figure 2.

These SCMs also include latent variables $\boldsymbol{U}$. According to the SCMs, the covariates $\boldsymbol{X}$ and the latent variables $\boldsymbol{U}$ are initially generated for the entire population. Subsequently, the true interference network $\boldsymbol{A}^*$ is generated, potentially depending on both $\boldsymbol{X}$ and $\boldsymbol{U}$. We consider two distinct scenarios for how the observed proxies $\mathcal{A}$ relate to $\boldsymbol{A}^*$.

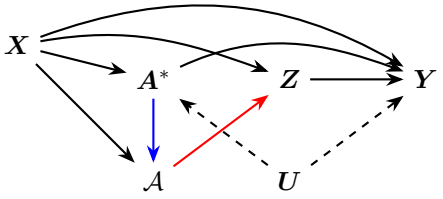
(1) **Causal proxies** (Figures 1(a)-(b)). Here, the proxies $\mathcal{A}$ are direct descendants of the true network $\boldsymbol{A}^*$. For example, such proxies represent noisy measurements of the true interference structure.

(2) **Non-causal proxies** (Figures 1(c)-(d)). In this scenario, both the true network $\boldsymbol{A}^*$ and the proxies $\mathcal{A}$ depend on additional latent variables $\boldsymbol{V}$, and no direct edges in the DAG link $\boldsymbol{A}^*$ and $\mathcal{A}$. For example, in a multilayer network (Kivelä et al., 2014), each layer represents a distinct type of relationship. The latent variables $\boldsymbol{V}$ may represent

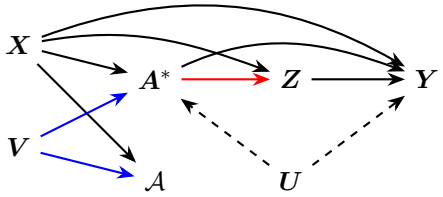
latent positions (Hoff et al., 2002), and the latent interference network $\boldsymbol{A}^*$ corresponds to one of these layers.



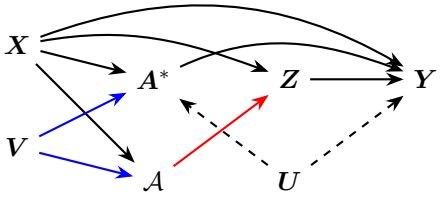
(a) Causal proxies & Latent assignment



(b) Causal proxies & Proxy assignment



(c) Non-causal proxies & Latent assignment



(d) Non-causal proxies & Proxy assignment

Figure 2: DAGs representing the assumed structural models for four different settings, corresponding to causal or non-causal proxies and treatment assignment based on the latent network $\boldsymbol{A}^*$ or the proxies $\mathcal{A}$: (a) Causal proxies with latent assignment. (b) Causal proxies with proxy assignment. (c) Non-causal proxies with latent assignment. (d) Non-causal proxies with proxy assignment. The dashed arrows are optional.

Next, we distinguish between two scenarios for the treatment assignment mechanism. In the first, treatments are assigned based on the latent network $\boldsymbol{A}^*$ (Figures 1(a) and (c)). This **latent assignment** is typical of observational studies, where researchers do not control the treatment assignment, and $\boldsymbol{Z}$ may be a function of the true network $\boldsymbol{A}^*$ rather than the proxies. In the second scenario, treatments are assigned based on the proxy networks $\mathcal{A}$ (Figures 1(b) and (d)). This **proxy assignment** is common in randomized trials, such as graph cluster randomization, where the network is clustered and treatments are randomized to clusters (Ugander et al., 2013; Eckles et al., 2017). As the true network $\boldsymbol{A}^*$ is latent, treatment allocation in these designs will often be a function of the observed proxies $\mathcal{A}$. In addition, we also consider scenarios where treatment assignment is influenced by both the true network $\boldsymbol{A}^*$ and the observed proxies $\mathcal{A}$. Such scenarios can occur in observational studies where both the latent and observed interactions influence the treatment assignment. In Appendix A, we present the DAGs and structural equations corresponding to this setting of combined treatment assignment under causal or non-causal proxies.

Each of the DAGs in Figure 2 implies a different set of structural equations for each of the four scenarios. We describe here the structural equations for the scenario of causal proxies with latent assignment (Figure 1(a)), and provide the set of equations corresponding to each

of the other DAGs in Appendix A. The structural equations corresponding to Figure 1(a) are

$$\begin{aligned}
\boldsymbol{U}_i &= f_U\big(\boldsymbol{\varepsilon}_{U_i}\big), && i \in [N] \\
\boldsymbol{X}_i &= f_X\big(\boldsymbol{\varepsilon}_{X_i}\big), && i \in [N] \\
\boldsymbol{A}^* &= f_{A^*}\big(\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{\varepsilon}_{A^*}\big), && \\
\boldsymbol{A}_b &= f_{A_b}\big(\boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\varepsilon}_{A_b}\big), && b \in [B] \\
Z_i &= f_Z\big(\boldsymbol{A}^*, \boldsymbol{X}_i, \boldsymbol{X}_{-i}, \varepsilon_{Z_i}\big), && i \in [N] \\
Y_i &= f_Y\big(Z_i, \boldsymbol{X}_i, \boldsymbol{Z}_{-i}, \boldsymbol{X}_{-i}, \boldsymbol{A}^*, \boldsymbol{U}_i, \varepsilon_{Y_i}\big) && i \in [N],
\end{aligned} \tag{1}$$

where $f_U, f_X, f_{A^*}, f_{A_b}, f_Z, f_Y$ are fixed functions and $\boldsymbol{\varepsilon}_U, \boldsymbol{\varepsilon}_X, \boldsymbol{\varepsilon}_{A^*}, \boldsymbol{\varepsilon}_{A_b}, \boldsymbol{\varepsilon}_Z, \boldsymbol{\varepsilon}_Y$ are noise terms. We assume the pairwise noise vectors are independent, e.g., $\boldsymbol{\varepsilon}_Z \perp\!\!\!\perp \boldsymbol{\varepsilon}_Y$, but do not limit the dependence structure between units (e.g., the dependence of ε_{Y_i} and ε_{Y_j}). The proxy models f_{A_b} can be adapted to include settings where proxies also influence each other. For example, if the proxies follow an auto-regressive model where $\boldsymbol{A}_b$ is a function of $\boldsymbol{A}^*$ and all preceding proxies $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{b-1}$, we can modify $\boldsymbol{A}_b$ model in (1) to

$$\begin{aligned}
\boldsymbol{A}_1 &= f_{A_1}\big(\boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\varepsilon}_{A_1}\big), \\
\boldsymbol{A}_b &= f_{A_b}\big(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{b-1}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\varepsilon}_{A_b}\big), \qquad b > 1.
\end{aligned}$$

Specific network models for $\boldsymbol{A}^*$ and proxies $\mathcal{A}$ are detailed in Section 3.2. In the non-causal proxies scenarios (Figures 1(c) and 1(d)), we add a model for the latent variables $\boldsymbol{V}_i = f_V(\boldsymbol{\varepsilon}_{V_i}), i \in [N]$, let $\boldsymbol{V}$ influence $\boldsymbol{A}^*$ by including $\boldsymbol{V}$ in f_{A^*}, and replace $\boldsymbol{A}^*$ with $\boldsymbol{V}$ in f_{A_b}. In the proxy-based treatment assignments (Figures 1(b) and 1(d)), we replace $\boldsymbol{A}^*$ in f_Z with the proxies $\mathcal{A}$. Details are given in Appendix A.

Under all the SCMs, the outcomes $\boldsymbol{Y}$ are a function of all preceding variables except $\mathcal{A}$. Specifically, $\boldsymbol{Z}$ influences the outcome of each unit Y_i through the direct treatment, Z_i, and the treatments of other units, $\boldsymbol{Z}_{-i}$. Structural assumptions regarding interference can be expressed by the specification of f_Y. For example, if interference is assumed to occur only between neighbors in $\boldsymbol{A}^*$, then f_Y depends on $\boldsymbol{Z}_{-i}$ only through $\boldsymbol{Z}_{\mathcal{N}_i^*}$, where $\mathcal{N}_i^* = \{j : \boldsymbol{A}_{ij}^* = 1\}$ is the neighbor set of unit i in $\boldsymbol{A}^*$. The outcome model in (1) includes high-dimensional variables: covariates of other units $\boldsymbol{X}_{-i} \in \mathbb{R}^{(N-1)\times q}$, treatments of other units, $\boldsymbol{Z}_{-i} \in \mathbb{R}^{N-1}$, and the true network $\boldsymbol{A}^* \in \{0,1\}^{N\times N}$. Practical applications often require simplifying this complex dependence through modeling assumptions. We assume that Y_i depends on these variables only through summarizing functions ϕ_1, ϕ_2, and ϕ_3, leading to a modified SCM for $\boldsymbol{Y}$

$$Y_i = f_Y\big(Z_i, \boldsymbol{X}_i, \phi_1(\boldsymbol{Z}_{-i}, \boldsymbol{A}^*), \phi_2(\boldsymbol{X}_{-i}, \boldsymbol{A}^*), \phi_{3,i}(\boldsymbol{A}^*), \boldsymbol{U}_i, \varepsilon_{Y_i}\big),\ i \in [N]. \tag{2}$$

The function $\phi_1(\boldsymbol{Z}_{-i}, \boldsymbol{A}^*)$ defines the *exposure mapping*, summarizing treatments assigned to units other than i. For example, under the common assumption of neighborhood interference (Forastiere et al., 2020; Ogburn et al., 2024), this function simplifies to $\phi_1(\boldsymbol{Z}_{\mathcal{N}_i^*}, \boldsymbol{A}^*)$, which depends solely on the treatments of immediate neighbors. For binary treatments, a typical example of such an exposure mapping is the proportion of treated neighbors, $\phi_1(\boldsymbol{Z}_{\mathcal{N}_i^*}, \boldsymbol{A}^*) = (d_i^*)^{-1} \sum_{j\in\mathcal{N}_i^*} Z_j$, where $d_i^* = |\mathcal{N}_i^*|$ is the degree of unit i. The neighborhood interference assumption can be relaxed, for example, by considering higher-order neighbors

with decreasing magnitudes (Leung, 2022). Heterogeneous effects can be included by defining appropriate $\phi_1 = \phi_1(\boldsymbol{Z}_{\mathcal{N}_i^*}, \boldsymbol{X}_{\mathcal{N}_i^*}, \boldsymbol{A}^*)$. The function $\phi_2(\boldsymbol{X}_{-i}, \boldsymbol{A}^*)$ typically summarizes covariates of neighbors, for example, by using the mean or sum of neighbors' covariates (Tchetgen Tchetgen et al., 2020; Ogburn et al., 2024). Finally, the function $\phi_3(\boldsymbol{A}^*)$ reduces the network $\boldsymbol{A}^*$ into summary statistics such as node centrality, degree, or eigenvectors; see Kolaczyk (2009) for additional examples and analysis of network summary statistics. Here, $\phi_{3,i}$ represents the statistics corresponding to unit i.

For Figures 1(a)-(c), all back-door paths between $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are blocked by conditioning on the covariates $\boldsymbol{X}$ and the true network $\boldsymbol{A}^*$. For Figure 1(d), conditioning on $\boldsymbol{X}$ and $\boldsymbol{A}^*$ alone is not sufficient because $\boldsymbol{A}^*$ becomes a collider on the back-door path $\boldsymbol{Y} \leftarrow \boldsymbol{U} \rightarrow \boldsymbol{A}^* \leftarrow \boldsymbol{V} \rightarrow \mathcal{A} \rightarrow \boldsymbol{Z}$. However, this back-door path can be blocked by the observed proxies $\mathcal{A}$, which implies that the conditioning set in this scenario is $\boldsymbol{X}, \boldsymbol{A}^*$, and $\mathcal{A}$. Our SCMs explicitly allow latent variables $\boldsymbol{U}$ to affect both the true interference network $\boldsymbol{A}^*$ and the outcomes $\boldsymbol{Y}$, but exclude direct effects of $\boldsymbol{U}$ on treatments $\boldsymbol{Z}$. Such latent confounding might arise due to *latent homophily* (Shalizi and Thomas, 2011), where unmeasured similarities among units influence both network formation and outcomes — a phenomenon also known as *network endogeneity* (Goldsmith-Pinkham and Imbens, 2013). As we discuss in Section 5, our Bayesian inferential framework can naturally accommodate this latent confounding.

We adopt the standard positivity assumption from graphical models (Pearl, 2009), requiring that all variables have positive joint probability across their respective supports, formally $p(\boldsymbol{Y}, \boldsymbol{Z}, \mathcal{A}, \boldsymbol{A}^*, \boldsymbol{V}, \boldsymbol{X}, \boldsymbol{U}) > 0$. This assumption can be relaxed to conditional positivity, requiring positivity to hold only given $\boldsymbol{V}, \boldsymbol{U}, \boldsymbol{X}$.

2.3 Causal Estimands

Our analysis focuses on estimands that quantify the effects of hypothetical interventions on the population-level treatments, corresponding to various treatment assignment policies. Although the definition of estimands does not depend on the exposure mapping ϕ_1, their numerical values rely on modeling assumptions. We explicitly distinguish between structural assumptions, such as those concerning the exposure mapping, and causal estimands, which represent interventions directly on the population treatments $\boldsymbol{Z}$ (Sävje, 2024). Interventions on exposure levels (defined by ϕ_1) are generally not feasible in practice, especially at the population level. Throughout the paper, we treat the population covariates $\boldsymbol{X}$ as fixed. All estimands we define are explicitly conditioned on the true latent network $\boldsymbol{A}^*$ and covariates $\boldsymbol{X}$, although we often omit this conditioning for brevity. Following the terminology in Li et al. (2023), the estimands can be viewed as mixed average treatment effects.

Let $Y_i(\boldsymbol{z})$ be the outcome of unit i under an intervention that sets the treatment vector $\boldsymbol{Z}$ to a specific value $\boldsymbol{z}$ with probability one, similar to the operator $do(\boldsymbol{Z} = \boldsymbol{z})$ (Pearl, 2009). Using the SCM notations in (1), $Y_i(\boldsymbol{z})$ corresponds to the outcome under the intervention that sets

$$\begin{aligned} Z_i &= z_i && i \in [N] \\ Y_i &= f_Y\big(z_i, \boldsymbol{X}_i, \boldsymbol{z}_{-i}, \boldsymbol{X}_{-i}, \boldsymbol{A}^*, \boldsymbol{U}_i, \varepsilon_{Y_i}\big) && i \in [N]. \end{aligned}$$

We consider estimands corresponding to three distinct types of treatment policies: *static*, *dynamic*, or *stochastic* (Pearl, 2009; Ogburn et al., 2024). Let $\mu_i(\boldsymbol{z}) \equiv \mu_i(\boldsymbol{z}; \boldsymbol{A}^*, \boldsymbol{X}) =$

$\mathbb{E}[Y_i(\boldsymbol{z}) \mid \boldsymbol{A}^*, \boldsymbol{X}]$ be the expected outcome for unit i under an intervention setting the treatment vector $\boldsymbol{Z}$ to $\boldsymbol{z}$, given the true network and covariates.

We define the *static policy* as setting the treatment vector $\boldsymbol{Z}$ to a fixed value $\boldsymbol{z}$. Let $\mu(\boldsymbol{z}) = N^{-1}\sum_{i=1}^{N}\mu_i(\boldsymbol{z})$. The corresponding estimand compares two static interventions

$$\tau(\boldsymbol{z}, \boldsymbol{z}') = \mu(\boldsymbol{z}) - \mu(\boldsymbol{z}').$$

For example, for binary treatments, $\tau(\mathbf{1}, \mathbf{0})$ is the TTE – the effect of treating all units versus treating none.

A *dynamic policy* assigns treatments deterministically based on covariates and the network structure. Let $h(\boldsymbol{X})$ denote a deterministic function that sets treatments based on covariates $\boldsymbol{X}$. With a slight abuse of notation, we write the corresponding average expected outcome as $\mu(h) = N^{-1}\sum_{i=1}^{N}\mu_i(h)$. The estimand

$$\tau(h_1, h_0) = \mu(h_1) - \mu(h_0),$$

compares two dynamic policies, h_1 and h_0. For example, comparing two different age thresholds when studying policies that treat all individuals above a certain age.

Finally, a *stochastic policy* assigns treatments according to a distribution $\pi_{\boldsymbol{\alpha}}(\boldsymbol{z})$, parameterized by $\boldsymbol{\alpha}$. The policy $\pi_{\boldsymbol{\alpha}}$ can depend on $\boldsymbol{X}$, but we omit this option for brevity. Under such a policy, the intervention $do(\boldsymbol{Z} = \boldsymbol{z})$ occurs with probability $\pi_{\boldsymbol{\alpha}}(\boldsymbol{z})$ (Pearl, 2009). The average of expected outcomes marginalized over the stochastic policy is

$$\mu(\boldsymbol{\alpha}) = N^{-1}\sum_{i=1}^{N}\sum_{\boldsymbol{z}}\pi_{\boldsymbol{\alpha}}(\boldsymbol{z})\mu_i(\boldsymbol{z}).$$

Accordingly, the estimand $\tau(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_0)$ evaluates the impact of implementing one stochastic policy $\pi_{\boldsymbol{\alpha}_1}$ compared to another $\pi_{\boldsymbol{\alpha}_0}$. For example, the effect of randomly treating 80% versus 20% of the population. The estimand $\tau(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_0)$ is similar to the overall causal effect defined by Hudgens and Halloran (2008). Compared to Hudgens and Halloran (2008), who considered a design-based framework, we replace the outcomes under the intervention with their expected value. Similar approaches were taken by others (Tchetgen Tchetgen et al., 2020; Ogburn et al., 2024).

Under the SCMs described in Figures 1(a)–(c), all back-door paths between $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are blocked by conditioning on $\boldsymbol{X}$ and $\boldsymbol{A}^*$. Thus, using the back-door adjustment formula (Pearl, 2009), we have

$$\mu_i(\boldsymbol{z}) = \mathbb{E}[Y_i \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{A}^*, \boldsymbol{X}]$$

Under the modified SCM (2), this is further simplified. In the case of Figure 1(d), where conditioning on $\boldsymbol{A}^*$ opens a back-door path between $\boldsymbol{Z}$ and $\boldsymbol{Y}$, an additional conditioning on the proxies $\mathcal{A}$ is required for identification. The back-door adjustment is therefore $\mu_i(\boldsymbol{z}) = \mathbb{E}[Y_i \mid \boldsymbol{Z} = \boldsymbol{z}, \mathcal{A}, \boldsymbol{A}^*, \boldsymbol{X}]$.

Since the interference network $\boldsymbol{A}^*$ and the parameters governing the conditional expectation of outcomes $\boldsymbol{Y}$ are both latent, all estimands depend on these unknowns. The interpretation of the estimands is therefore conditional on the true interference network $\boldsymbol{A}^*$, similar to Ogburn et al. (2024), with the distinction that here $\boldsymbol{A}^*$ is latent. Thus, the interpretation of the estimands is the same had the network been known. Note also

that because we consider interventions on the population treatments, the interpretation of the causal estimands does not need to take the network into account. For example, the interpretation of randomly treating 80% versus 20% of the population remains the same. In Section 5.3, we describe how to estimate the estimands using samples from the posterior distribution.

2.4 Probabilistic Models

While the SCMs described above provide a general causal framework, practical applications require specifying probabilistic models for some of the structural equations. The components requiring modeling depend on the assumed SCM structure. For example, whether the proxies are causal or not and whether treatment assignment is based on the latent network or observed proxies. Throughout our analysis, we treat the covariates $\boldsymbol{X}$ as fixed.

Consider the setting of causal proxies with latent assignment (Figure 1(a)). In this case, researchers must specify probabilistic models for the outcome, treatment assignment, latent network, and proxy networks measurement mechanisms. Let $\boldsymbol{\eta}$ parameterize the outcome model $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta})$, $\boldsymbol{\beta}$ parameterize the treatment assignment model $p(\boldsymbol{Z} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\beta})$, $\boldsymbol{\theta}$ the true network model $p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta})$, and $\boldsymbol{\gamma}$ the observed proxy networks model $p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma})$. Let $\boldsymbol{\Theta} = (\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ denote the set of all the structural model parameters. Under this specification, the complete data likelihood of $(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^*)$ can be factorized according to Figure 1(a) as

$$p(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta}) = p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta}) p(\boldsymbol{Z} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\beta}) p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma}) p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta}), \tag{3}$$

highlighting that the complete data likelihood is composed of three data modules (outcomes, treatments, and proxies) and the prior network model.

In scenarios with proxy-based treatment assignment (Figures 1(b) and (d)), the treatment assignment model $p(\boldsymbol{Z} \mid \mathcal{A}, \boldsymbol{X})$ may depend on the proxies $\mathcal{A}$ but not on $\boldsymbol{A}^*$. In scenarios with combined treatment assignment (Appendix A), where both $\boldsymbol{A}^*$ and $\mathcal{A}$ influence $\boldsymbol{Z}$, the treatment assignment model is $p(\boldsymbol{Z} \mid \mathcal{A}, \boldsymbol{A}^*, \boldsymbol{X})$. Similarly, in non-causal proxy settings (Figures 1(c)-(d)), the proxy model $p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma})$ and true network model $p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta})$ are replaced by models conditioned on the latent common causes $\boldsymbol{V}$, i.e., $p(\mathcal{A} \mid \boldsymbol{V}, \boldsymbol{X}, \boldsymbol{\gamma})$ and $p(\boldsymbol{A}^* \mid \boldsymbol{V}, \boldsymbol{X}, \boldsymbol{\theta})$. Scenarios with latent variables $\boldsymbol{U}$ are discussed in Appendix C.

3 Proxy Measurements and Network Models

3.1 Types of Proxy Networks

We categorize the observed proxy measurements $\mathcal{A}$ of the true latent interference network $\boldsymbol{A}^*$ into the following types:

(i) **Measurement error.** Researchers observe a single (Butts, 2003) or multiple (De Bacco et al., 2023) proxy networks, each reflecting the true network $\boldsymbol{A}^*$ with varying degrees of error, which can depend on covariates. For unweighted networks, measurement error corresponds to edge misclassification. For instance, in survey-derived networks, respondents might forget certain connections or misunderstand reporting instructions, resulting in missing or spurious edges.

Another scenario occurs under partial interference assumptions (Hudgens and Halloran, 2008), where researchers observe networks composed of distinct clusters, while the true interference network may have cross-cluster relations, thus ignoring potential cross-cluster interactions (Weinstein and Nevo, 2026).

(ii) **Multiple sources.** Researchers obtain several proxies derived from distinct data sources, each potentially providing complementary information about the latent interference network. These sources can include online platforms, surveys, biological measurements, physical traces, or spatial data, each capturing different relational aspects. For example, Goyal et al. (2023) combined epidemiological contact tracing, genetic testing, and behavioral surveys to estimate disease transmission networks. Although valuable, integrating multiple proxy networks poses inference challenges, as proxies may differ in their accuracy and relevance for capturing the interference mechanism under study

(iii) **Multilayer networks.** Researchers collect multiple networks representing distinct types of relationships among the same set of units (Kivelä et al., 2014). Unlike multiple-source proxies that offer complementary information about a single underlying structure, multilayer networks explicitly represent fundamentally different social relations. An example would be one network layer representing friendship ties and another representing work collaboration. Within our framework, multilayer networks can be represented using two modeling approaches. In the first, the observed layers are non-causal proxies (Figures 1(c)-(d)), generated from shared latent positions (Salter-Townshend and McCormick, 2017; Sosa and Betancourt, 2022). Alternatively, one can view the true latent network $\boldsymbol{A}^*$ as a common ancestor of all the observed layers, representing causal proxies (Figures 1(a)-(b)).

Next, we show how these different types of proxies can be expressed mathematically within our modeling framework.

3.2 True and Proxy Network Models

Statistical analysis of network data encompasses a wide variety of models suited for different applications (Fienberg, 2012). Common examples include random graphs (Erdős and Rényi, 1959), stochastic block models (SBM) (Holland et al., 1983), exponential random graphs (Holland and Leinhardt, 1981), and latent space models (LSM) (Hoff et al., 2002). Our framework is flexible regarding the choice of the model for the true network $\boldsymbol{A}^*$, subject to two practical constraints: computational feasibility and parameter identifiability. We further discuss the parametric identifiability of the structural model parameters $\boldsymbol{\Theta}$ in Section 4.

The observed proxy network(s) can be modeled to reflect different measurement scenarios. We now illustrate how some of the previously mentioned examples (Section 3.1) can be formalized within our framework.

Example 1 (Random noise)**.** A single network $\boldsymbol{A}$ is observed, generated by independently measuring each edge of the latent network $\boldsymbol{A}^*$ with random error $\Pr(A_{ij} = 1 \mid A^*_{ij} = k, \boldsymbol{\gamma}) = \gamma_k,\ k = 0, 1$, corresponding to false positive rate when $k = 0$ and true positive rate when

$k = 1$. Extending to covariate-dependent measurement error is possible by modeling the error rates as a function of covariates $\boldsymbol{X}$.

Example 2 (Censoring)**.** The observed network $\boldsymbol{A}$ is a censored version of $\boldsymbol{A}^*$, namely, each unit is allowed to report at most $C > 0$ edges (friends) from its true connections (Griffith, 2021). If unit i has more than C true edges (i.e., if $d_i^* > C$ where d_i^* is its degree in $\boldsymbol{A}^*$), assume the unit independently reports a random subset of C neighbors; otherwise, it reports all. Assuming an edge is observed if either node reports it and no false positives, i.e., $\Pr(A_{ij} = 1 \mid A_{ij}^* = 0) = 0$, the reporting probability for existing edges is

$$\Pr(A_{ij} = 1 \mid A_{ij}^* = 1) = 1 - \left[1 - \min\left(1, \frac{C}{d_i^*}\right)\right] \times \left[1 - \min\left(1, \frac{C}{d_j^*}\right)\right].$$

Example 3 (Cross-cluster contamination)**.** A single network $\boldsymbol{A}$ consisting of well-separated clusters is observed, while $\boldsymbol{A}^*$ is generated from a network model that may include cross-cluster edges, such as SBM (Holland et al., 1983). Researchers assume that interference is only possible within clusters (partial interference), but cross-cluster contamination exists. Let $X_{1,i}$ denote the cluster of unit i. One possible measurement model is

$$\Pr(A_{ij} = 1 \mid A_{ij}^* = k, X_{1,i}, X_{1,j}, \boldsymbol{\gamma}) = \gamma_k \mathbb{I}\{X_{1,i} = X_{1,j}\},\ k = 0, 1,$$

with false positive rate γ_0 and true positive rate γ_1 within clusters, and no edges observed between clusters. This model can be generalized, e.g., by replacing γ_k with $\gamma_{k,ij}$ that depends on the cluster memberships. While the proxy network $\boldsymbol{A}$ provides no information regarding cross-cluster edges, our estimation framework allows for learning these structures through other available modules, as described in Section 5.

Example 4 (Repeated noisy measurements)**.** Multiple proxy networks are observed ($B > 1$), each is a noisy measurement of the latent network $\boldsymbol{A}^*$ (De Bacco et al., 2023). Each proxy follows some measurement model $p(\boldsymbol{A}_b \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma}_b)$ which may be identical or vary across proxies. Dependence among proxies can be modeled hierarchically or temporally. In a temporal setting, the measurement of network $\boldsymbol{A}_b$ can follow an auto-regressive model, e.g., by taking $p(\boldsymbol{A}_b \mid \boldsymbol{A}_1, \ldots, \boldsymbol{A}_{b-1}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma}_b)$.

Example 5 (Multilayer networks)**.** Researchers observe multiple networks on the same N units, each representing distinct relationship types. As previously discussed, these observed multilayer networks can be modeled under two different assumptions about their relationship with the latent network $\boldsymbol{A}^*$:

(a) **Non-causal proxies.** Both the true network $\boldsymbol{A}^*$ and the observed proxies $\mathcal{A}$ depend on common latent variables $\boldsymbol{V}$ (Figure 1(a)-(b)). Given these latent variables and covariates $\boldsymbol{X}$, the observed proxies and the latent network are independent. For instance, each network layer (including $\boldsymbol{A}^*$) can be generated from a latent space model (LSM; Hoff et al., 2002), where $\boldsymbol{V}$ represents latent unit-level positions. Latent positions $\boldsymbol{V}$ can either be shared across all layers, including the latent network (Salter-Townshend and McCormick, 2017), or structured hierarchically (Sosa and Betancourt, 2022). Under this setup, proxy networks $\mathcal{A}$ provide information about $\boldsymbol{A}^*$ only indirectly through their common latent positions $\boldsymbol{V}$.

(b) **Causal proxies.** The network $\boldsymbol{A}^*$ is a latent predecessor of all the observed proxies $\mathcal{A}$ (Figure 1(c)-(d)). For example, each proxy network layer b can be modeled as

$$\Pr_b(A_{b,ij} = 1 \mid A^*_{ij}, \boldsymbol{X}, \boldsymbol{\gamma}_b) = r^{-1}\big(\gamma_{b,0} A^*_{ij} + \boldsymbol{\gamma}_b' \widetilde{\boldsymbol{X}}_{ij}\big),$$

where $r : [0,1] \to \mathbb{R}$ is a link function and $\widetilde{\boldsymbol{X}}_{ij}$ are edge-level covariates, derived from the nodal covariates $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ (e.g., as done in Section 6).

4 Identifiability of Structural Parameters

Under the proposed class of SCMs (Figure 2), establishing the identifiability of the structural parameters $\boldsymbol{\Theta}$ is a fundamental prerequisite for valid statistical inference. It ensures that the true structural parameters $\boldsymbol{\Theta}$ can, in principle, be uniquely recovered from the distribution of the observed data $\boldsymbol{\mathcal{D}} = (\boldsymbol{Y}, \boldsymbol{Z}, \mathcal{A}, \boldsymbol{X})$. In the absence of identifiability, the posterior distribution in a Bayesian framework (Section 5) may be heavily influenced by the prior, as the data alone are insufficient to uniquely recover the parameters.

In our framework, the network $\boldsymbol{A}^*$ is latent, rendering the observed data likelihood a complex mixture model of the complete data likelihood (3), $p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\Theta}) = \sum_{\boldsymbol{A}^*} p(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta})$. Since this summation is computationally intractable due to the high-dimensional discrete space of $\boldsymbol{A}^*$, establishing global identifiability analytically for the general case is prohibitively complex.

Instead, we investigate the global identifiability of the structural parameters in simplified settings. We illustrate how the conditional independence structure implied by the SCMs allows for the unique recovery of parameters using moment equations (Rothenberg, 1971). Specifically, because the outcomes and proxy networks are conditionally independent given the latent network $\boldsymbol{A}^*$, their joint moments provide distinct moment equations that can be solved to recover the parameters $\boldsymbol{\Theta}$. This analysis highlights the mechanism through which the distinct data modules inform the structural parameters. This perspective aligns with the identification results for latent variable models using at least three independent views (Allman et al., 2009).

In addition, we discuss sufficient conditions for local identifiability, which require that the likelihood function is not flat in a neighborhood of the true parameters. This is assessed via the nonsingularity of the Fisher Information Matrix (FIM). In Appendix B, we utilize the Missing Information Principle (Louis, 1982) to characterize the observed data FIM as the difference between the complete data information and the missing information induced by the latent network. Intuitively, we show that the parameters are locally identifiable if the information provided by the distinct data modules dominates the uncertainty surrounding the latent network. Details are provided in Appendix B.

Remark 1 (From Structural Parameters to Causal Estimands)**.** *The causal estimands defined in Section 2.3 are functions of the specific realization of the latent network $\boldsymbol{A}^*$. Consequently, establishing the identifiability of $\boldsymbol{\Theta}$ does not imply that the causal effects are nonparametrically identified, as $\boldsymbol{A}^*$ remains latent. Instead, the identification of $\boldsymbol{\Theta}$ allows for the recovery of the conditional distribution of the latent network given the data, $p(\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta})$, and subsequently, the distribution of the causal estimands. This distinction*

is central to our framework: we do not seek to recover a point estimate of the causal effects, but rather to recover their posterior distribution. This motivates our Bayesian inference approach in Section 5, which targets this full posterior distribution.

4.1 Global Identifiability In Simplified Settings

Global identifiability requires that the map from the structural parameters $\boldsymbol{\Theta}$ to the observed data likelihood $p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\Theta})$ is injective. Define the *moment map* $\mathcal{M}(\boldsymbol{\Theta}) = \mathbb{E}_{\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}}[\mathbf{m}(\boldsymbol{\mathcal{D}})]$ as the expected value of a vector of observable moments $\mathbf{m}(\boldsymbol{\mathcal{D}})$. The parameter $\boldsymbol{\Theta}$ is globally identified with respect to these moments if $\mathcal{M}(\boldsymbol{\Theta})$ is injective (Rothenberg, 1971).

In the context of the proposed SCMs, we construct these moment equations by exploiting the conditional independence structure of the model. For instance, in Figure 2(a), the independence of outcomes $\boldsymbol{Y}$ and proxies $\mathcal{A}$ given the latent network $\boldsymbol{A}^*$ and covariates $\boldsymbol{X}$ implies that the observed correlation (given $\boldsymbol{X}$) between $\boldsymbol{Y}$ and $\mathcal{A}$ is entirely mediated by $\boldsymbol{A}^*$. Consequently, the observable cross-moments become explicit functions of $\boldsymbol{\Theta}$. By deriving these analytical expectations, we can construct a system of moment equations and assess the injectivity of the map. We now demonstrate this identification strategy by explicitly deriving the system of moment equations for a simplified class of models.

Consider a setup without covariates, where the true network $\boldsymbol{A}^*$ follows an Erdős-Rényi model (Erdős and Rényi, 1959), the outcome model is linear with exposure mapping ϕ_1 (as in (2)) representing the number of treated neighbors, and proxy network measurements are subject to random measurement error as in Example 1. The models are specified as follows:

$$\begin{aligned} A^*_{ij} &\sim \text{Ber}(\theta), \\ A_{ij} \mid A^*_{ij} = k &\sim \text{Ber}(\gamma_k),\ k = 0, 1, \\ Z_i &\sim \text{Ber}(p_z), \\ Y_i &= \eta_1 Z_i + \eta_2 \sum_{j \neq i} A^*_{ij} Z_j + \varepsilon_i, \end{aligned} \tag{4}$$

where ε_i are independent and zero-mean error terms. The set of parameters is $\boldsymbol{\Theta} = (\theta, \gamma_0, \gamma_1, p_z, \eta_1, \eta_2)$. The parameters p_z and η_1 are identified from the observed first moments $\mathbb{E}[Z_i]$ and $\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]$, respectively. Deriving the analytical moments (Appendix B), we obtain the following system of three equations:

$$\begin{aligned} \mathbb{E}[Y_i] &= \eta_1 p_z + \eta_2 (N-1) p_z \theta, \\ \Pr(A_{ij} = 1) &= \gamma_1 \theta + \gamma_0 (1 - \theta), \\ \text{Cov}(Y_i, d_i^{obs}) &= \eta_2 (N-1) p_z \theta (1-\theta)(\gamma_1 - \gamma_0), \end{aligned}$$

where $d_i^{obs} = \sum_{j \neq i} A_{ij}$ is the observed degree of unit i in the proxy network. This system presents three equations for four unknowns $(\theta, \gamma_0, \gamma_1, \eta_2)$. Consequently, the simplified model (4) with a single proxy network is not globally identifiable without further restrictions. However, this lack of identification is resolved under several realistic extensions of the model, which provide the necessary additional information to achieve global identifiability (see Appendix B for derivations):

1. **Constraints on Measurement Error:** If reliability data is available such that the false positive rate γ_0 is exogenously identified (or fixed to zero), the system reduces to three unknowns, rendering the parameters globally identifiable.

2. **Multiple Proxies:** When a second proxy network $\boldsymbol{A}_2$ is available, such that $\boldsymbol{A}_2 \perp\!\!\!\perp \boldsymbol{A}_1 | \boldsymbol{A}^*$, and $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are identically distributed, we can utilize the expected disagreement between the proxies $\frac{2}{N(N-1)} \sum_{i<j} |A_{1,ij} - A_{2,ij}|$. This additional moment allows for the unique recovery of all four parameters $(\theta, \gamma_0, \gamma_1, \eta_2)$.

3. **Network-Dependent Treatment:** Consider settings where treatment assignment depends on the latent network (Figures 1(a),(c)). For example, the treatment assignment mechanism is $\Pr(Z_i = 1 \mid \boldsymbol{A}^*) \propto d_i^*$, where d_i^* is the degree of i in the true network. The observed treatments act as an additional proxy source. Treatments are independent only conditional on the latent network, thus $\mathbb{E}[Z_i Z_j]$ provides an additional equation involving θ and p_z. Also, the covariance $\mathrm{Cov}(Z_i, d_i^{obs})$ provides an additional equation linking θ and the proxy error rates. Thus, we have over-identification of the structural parameters.

4. **Network-Correlated Outcomes:** In many applications, the outcome errors ε_i are correlated among neighbors. For example, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \rho A_{ij}^*$. While this introduces a new parameter ρ, it allows us to leverage the second moments of the outcome. Specifically, the conditional expectations $\mathbb{E}[Y_i Y_j \mid A_{ij} = k]$ for $k \in \{0, 1\}$ provides two distinct equations that allow for the unique identification of the full parameter set, including ρ.

These results highlight a fundamental insight: while the latent nature of $\boldsymbol{A}^*$ poses a challenge, the parameters $\boldsymbol{\Theta}$ are recoverable provided there is sufficient information from the outcomes, treatments, and proxy networks. Furthermore, under (4), had we ignored the outcomes and treatments and tried to identify $(\gamma_0, \gamma_1, \theta)$ from the proxy networks $\mathcal{A}$ alone, we would have needed at least three independent proxy networks to achieve identifiability (Chang et al., 2022). That illustrates that in our framework, each data module (e.g., outcomes, treatments, proxies) contributes to identifying the structural parameters. That is, θ and γ are recovered not just using the proxies $\mathcal{A}$, but also through information contained in the outcomes $\boldsymbol{Y}$ and possibly the treatments $\boldsymbol{Z}$ modules.

5 Bayesian Estimation and Inference

We propose a Bayesian framework for estimation. We describe here inference in the scenario without latent variables $\boldsymbol{U}$ and with causal proxies $\mathcal{A}$ (Figures 1(a)-(b)). The details for the scenarios involving latent variables $\boldsymbol{U}$ or non-causal proxies (Figures 1(c)-(d)) are given in Appendix C.

We first consider the setting of treatment assignment based on the latent network (Figure 1(a)). We assume prior independence $p(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = p(\boldsymbol{\eta}) p(\boldsymbol{\beta}) p(\boldsymbol{\theta}) p(\boldsymbol{\gamma})$. Recall that the observed data are $\boldsymbol{\mathcal{D}} = (\boldsymbol{Y}, \boldsymbol{Z}, \mathcal{A}, \boldsymbol{X})$. The latent variables are the true network $\boldsymbol{A}^*$ and the parameters $\boldsymbol{\Theta} = (\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma})$. The factorization (3) and prior independence yield the joint

posterior

$$
\begin{aligned}
p(\boldsymbol{\Theta}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}) \propto\ & p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta}) p(\boldsymbol{\eta}) \\
& \times p(\boldsymbol{Z} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\beta}) p(\boldsymbol{\beta}) \\
& \times p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma}) \\
& \times p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta}).
\end{aligned} \tag{5}
$$

When treatment assignment is based on the proxy networks (Figure 1(b)), the treatment model becomes $p(\boldsymbol{Z} \mid \mathcal{A}, \boldsymbol{X}, \boldsymbol{\beta})$. Consequently, by replacing $p(\boldsymbol{Z} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\beta})$ with $p(\boldsymbol{Z} \mid \mathcal{A}, \boldsymbol{X}, \boldsymbol{\beta})$ in (5), we can marginalize the posterior $p(\boldsymbol{\Theta}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}})$ over $\boldsymbol{\beta}$, yielding

$$
\begin{aligned}
p(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}) \propto\ & \int_{\boldsymbol{\beta}} p(\boldsymbol{\Theta}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}) d\boldsymbol{\beta} \\
\propto\ & p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta}) p(\boldsymbol{\eta}) \\
& \times p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma}) \\
& \times p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta}).
\end{aligned} \tag{6}
$$

In either scenario (5) or (6), augmenting the outcome model with propensity scores as additional covariates is possible but often requires the cutting of 'model feedback' (Zigler et al., 2013). We describe such extensions in Appendix C.

Sampling from the joint posterior (5) or (6) is challenging due to its mixed space, which includes continuous parameters $\boldsymbol{\Theta}$ and discrete latent variables $\boldsymbol{A}^*$. The discrete space grows super-exponentially with N, containing $\mathcal{O}(2^{N^2})$ terms. A common approach for mixed-space posteriors is to marginalize discrete variables (Stan Development Team, 2024). However, in our setting, the outcome model for each unit depends on $\boldsymbol{A}^*$ through its entire neighborhood $\boldsymbol{A}_i^*$ (and possibly beyond), inducing complex interdependencies that preclude obtaining a closed-form marginalized posterior (indeed, even for the simple linear model in (4), the marginal likelihood is analytically intractable). Brute-force marginalization of $\boldsymbol{A}^*$ requires summing over $\mathcal{O}\big(2^{N^2}\big)$ terms, making it computationally infeasible.

Given this intractability, we explored various alternatives. One approach is to relax the discrete space into a continuous one using reparameterization techniques (Zhang et al., 2012; Maddison et al., 2017; Nishimura et al., 2020). However, we found that in our setting, such continuous relaxation methods introduce approximation errors that result in unsatisfactory performance (Section 6). Another option is to use Mixed Hamiltonian Monte Carlo (HMC), which explicitly accounts for the discrete space (Zhou, 2020). We found that Mixed-HMC does not scale well in terms of computation time, even for small N. These limitations highlight the need for more efficient sampling strategies, which we explore next.

To this end, we propose a Block Gibbs MCMC algorithm that decomposes the sampling process into two blocks consisting of continuous parameters $\boldsymbol{\Theta}$ and discrete network $\boldsymbol{A}^*$. Given the network state $\boldsymbol{A}^*$, updating the continuous parameters $\boldsymbol{\Theta}$ is straightforward using standard MCMC techniques. The challenge lies in updating the latent network $\boldsymbol{A}^*$. Due to its large discrete spaces, updating $\boldsymbol{A}^*$ entries with a Metropolis-Hastings algorithm using a random walk Markov kernel can lead to poor mixing and slow convergence. Therefore, we use a variant of Locally Informed Proposals (LIP; Zanella, 2020). LIP uses information from the conditional posterior of $\boldsymbol{A}^*$ to explore possible updates efficiently and overcome the poor mixing of standard random walk proposals in high-dimensional discrete spaces.

We now provide details about LIP in our settings and then describe the complete Block Gibbs algorithm.

5.1 Locally Informed Proposals

In our case of a latent network with binary edges, each update of the current $\boldsymbol{A}^*$ state corresponds to 'flipping' some entries ($A^*_{ij} = 1 \to A^*_{ij} = 0$ or $A^*_{ij} = 0 \to A^*_{ij} = 1$). A standard random walk kernel proposes entries to be flipped at random. In high-dimensional spaces, such uninformed proposals may lead to states with low posterior probability, resulting in high rejection rates and slow mixing. To address this, we adopt the LIP framework, which constructs a proposal distribution that biases flips towards states with higher posterior probability. Theoretically, LIPs are designed to be locally reversible with respect to the target distribution, a property shown to improve the asymptotic efficiency and mixing times of the Markov chain (Zanella, 2020).

However, implementing exact LIP requires evaluating the posterior probability for every possible proposal of a single-edge flip. For a network of size N, this entails $\mathcal{O}(N^2)$ posterior evaluations per iteration, which can be computationally prohibitive. To render LIP feasible, we leverage the method proposed by Grathwohl et al. (2021), which approximates the differences using gradients. This reduces the computational cost of generating a proposal to a single gradient evaluation per iteration.

Consider the joint posterior (5). Given the other unknowns $\boldsymbol{\Theta}$ and the data $\boldsymbol{\mathcal{D}}$, the posterior of $\boldsymbol{A}^*$ is proportional to

$$p(\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta}) \propto \underbrace{p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta})}_{\text{Outcomes}} \times \underbrace{p(\boldsymbol{Z} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\beta})}_{\text{Treatments}} \times \underbrace{p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma})}_{\text{Proxies}} \times \underbrace{p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta})}_{\text{Prior}}. \tag{7}$$

Equation (7) illustrates that the conditional posterior of $\boldsymbol{A}^*$ is composed of the likelihood terms from all data modules dependent on $\boldsymbol{A}^*$ (outcomes, treatments, and proxies in this scenario), alongside the prior network model. Since the LIP algorithm utilizes this conditional posterior to propose updates for the Markov chain, the reconstruction of the latent network is not driven solely by the observed proxies $\mathcal{A}$. Instead, the inference actively leverages the signal embedded in the outcomes and, where applicable, treatments. In the specific case of posterior (6), treatment allocation depends solely on $\mathcal{A}$ and $\boldsymbol{X}$. Consequently, the treatment assignment model is constant with respect to $\boldsymbol{A}^*$, and the conditional posterior $p(\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta})$ is driven only by the outcomes, proxies, and the prior network model.

Let $p(\boldsymbol{A}^* \mid \cdot) \equiv p(\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta})$ denote the conditional posterior in (7). Since the interference network is undirected, we only need to update the upper (or lower) triangle values of $\boldsymbol{A}^*$ and set the lower (or upper) triangle values to be the same. The LIP for updating network state from $\boldsymbol{A}^*_t$ to $\boldsymbol{A}^*_{t+1}$ have the structure (Zanella, 2020)

$$Q(\boldsymbol{A}^*_{t+1} \mid \boldsymbol{A}^*_t, \cdot) \propto g\left(\frac{p(\boldsymbol{A}^*_{t+1} \mid \cdot)}{p(\boldsymbol{A}^*_t \mid \cdot)}\right) \mathbb{I}\big(\boldsymbol{A}^*_{t+1} \in H(\boldsymbol{A}^*_t)\big), \tag{8}$$

where $H(\boldsymbol{A}^*_t) = \big\{\boldsymbol{A}^* : \sum_{i<j} |A^*_{ij} - A^*_{t,ij}| = 1\big\}$ is the Hamming ball of size one, with $A^*_{t,ij}$ being the ij entry of $\boldsymbol{A}^*_t$, and g is a function that must satisfy the balancing condition $g(a) = ag(1/a)$ for all $a > 0$ for the proposal $Q(\cdot|\cdot)$ to be locally reversible with respect to

$p(\boldsymbol{A}^* \mid \cdot)$ (Zanella, 2020). Common choices include $g(a) = \sqrt{a}$ or $g(a) = \frac{a}{1+a}$. The proposals in (8) use the posterior ratio to inform possible moves constrained to flipping only one of $\boldsymbol{A}_t^*$ entries, and are typically followed by an accept/reject step.

Let $\log p(\boldsymbol{A}^* \mid \cdot)$ be the log of the conditional posterior in (7) and denote the difference between two subsequent states by $\Delta(\boldsymbol{A}_{t+1}^*, \boldsymbol{A}_t^* \mid \cdot) = \log p(\boldsymbol{A}_{t+1}^* \mid \cdot) - \log p(\boldsymbol{A}_t^* \mid \cdot)$. Assume that $g(a) = \sqrt{a}$. The proposals (8) can be written as

$$Q(\boldsymbol{A}_{t+1}^* \mid \boldsymbol{A}_t^*, \cdot) \propto \exp\Big(\frac{1}{2}\Delta(\boldsymbol{A}_{t+1}^*, \boldsymbol{A}_t^* \mid \cdot)\Big)\mathbb{I}\big(\boldsymbol{A}_{t+1}^* \in H(\boldsymbol{A}_t^*)\big). \tag{9}$$

Let $\boldsymbol{A}_{t+1}^*(ij)$ be $\boldsymbol{A}_t^*$ with only entry $A_{t,ij}^*$ flipped. With a slight abuse of notation, denote $\Delta(ij, t \mid \cdot) \equiv \Delta(\boldsymbol{A}_{t+1}^*(ij), \boldsymbol{A}_t^* \mid \cdot)$. Since $\boldsymbol{A}_{t+1}^*(ij) \in H(\boldsymbol{A}_t^*)$, proposing an update using (9) is equivalent to sampling an entry with probability $q(ij \mid \boldsymbol{A}_t^*, \cdot) \equiv \text{Softmax}\Big(\frac{1}{2}\Delta(ij, t \mid \cdot)\Big)$ and flip its value. However, that entails computing $\log p(\boldsymbol{A}_{t+1}^*(ij) \mid \cdot)$ for all $1 \leq i < j \leq N$ which entails $\mathcal{O}(N^2)$ evaluations of the log-posterior.

To overcome this computational challenge, we approximate the differences $\Delta(ij, t \mid \cdot)$ using gradients of the log-posterior. The key insight is that although $\boldsymbol{A}^*$ is discrete, the structure of the log-posterior often allows for accurate gradient-based approximations (Grathwohl et al., 2021). This structure enables us to approximate the differences Δ with a first-order Taylor series

$$\widetilde{\Delta}(ij, t \mid \cdot) = -(2A_{t,ij}^* - 1)\frac{\partial \log p(\boldsymbol{A}_t^* \mid \cdot)}{\partial A_{t,ij}^*} \approx \Delta(ij, t \mid \cdot). \tag{10}$$

In vector form, we can compute the gradients $\nabla \log p(\boldsymbol{A}_t^* \mid \cdot)$ with respect to entries in the upper triangle of $\boldsymbol{A}^*$ and approximate all the $N(N-1)/2$ differences $\Delta(ij, t \mid \cdot)$ through

$$\widetilde{\boldsymbol{\Delta}}(\boldsymbol{A}_t^* \mid \cdot) = -(2\text{vec}(\boldsymbol{A}_t^*) - 1) \odot \nabla \log p(\boldsymbol{A}_t^* \mid \cdot). \tag{11}$$

where $\odot$ is the element-wise product and $\text{vec}(\boldsymbol{A}^*)$ is the vector of $N(N-1)/2$ upper triangle values of $\boldsymbol{A}^*$. Thus, $\widetilde{\boldsymbol{\Delta}}(\boldsymbol{A}_t^* \mid \cdot) = \big\{\widetilde{\Delta}(ij, t) : 1 \leq i < j \leq N\big\}$ is the set of all approximate differences (10). Computing (11) requires only a single gradient evaluation and can be performed with any automatic differentiation library. We formally analyze the approximation error in Appendix C.5. We demonstrate that for an extended version of the simplified model (4), the efficiency of the gradient-based approximation relative to the exact proposal decreases as the interference signal-to-noise ratio increases, but, crucially, remains independent of the network size N. Furthermore, this approximation was accurate in the numerical illustrations (Appendix D).

The pseudocode for a single update of $\boldsymbol{A}^*$ state with LIP is given in Algorithm 1. The pseudocode is applicable for both posterior (5) and (6) by replacing $\log p(\cdot)$ with the corresponding log-posterior. We extend the single-entry update by proposing $K \geq 1$ flips per iteration. Instead of selecting a single entry, we sample K entries without replacement using the Gumbel-Max trick (Step 3). The selected indices are saved in a set $\mathcal{I}$. The function `FlipEntries` simply takes a network state $\boldsymbol{A}_t^*$ and flips the values of all entries in the set of indices $\mathcal{I}$. For $K = 1$, the algorithm performs *exact* LIP but is likely to require more iterations, while for $K > 1$, the updates become *approximate* but can accelerate convergence. Increasing K poses a trade-off between accuracy and computational efficiency. In our numerical illustrations (Section 6), we varied K depending on the network model and sample size.

Algorithm 1 A Single $\boldsymbol{A}^*$ Update with Locally Informed Proposals

Input: Data $\boldsymbol{\mathcal{D}}$, continuous parameters $\boldsymbol{\Theta}$, current state $\boldsymbol{A}_t^*$, log-posterior $\log p(\cdot)$, number of entries to update $K \geq 1$.

1: Compute differences $\widetilde{\boldsymbol{\Delta}}(\boldsymbol{A}_t^* \mid \cdot) \equiv \widetilde{\boldsymbol{\Delta}}(\boldsymbol{A}_t^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta})$ using (11).

2: Compute $q(ij \mid \boldsymbol{A}_t^*, \cdot) = \text{Softmax}\Big(\frac{1}{2}\widetilde{\boldsymbol{\Delta}}(\boldsymbol{A}_t^* \mid \cdot)\Big)$.

3: Sample without replacement K edges $\mathcal{I}$ using Gumbel-Max trick with probabilities $q(ij \mid \boldsymbol{A}_t^*, \cdot)$.

4: Compute *forward* probability $q(\mathcal{I} \mid \boldsymbol{A}_t^*, \cdot) = \prod_{ij \in \mathcal{I}} q(ij \mid \boldsymbol{A}_t^*, \cdot)$.

5: $\boldsymbol{A}_{t+1}^* \leftarrow \texttt{FlipEntries}(\boldsymbol{A}_t^*, \mathcal{I})$.

6: Compute *backward* probability $q(\mathcal{I} \mid \boldsymbol{A}_{t+1}^*, \cdot) = \prod_{ij \in \mathcal{I}} q(ij \mid \boldsymbol{A}_{t+1}^*, \cdot)$ similarly to step 2.

7: Accept with probability

$$\min\left(\exp\left(\Delta(\boldsymbol{A}_{t+1}^*, \boldsymbol{A}_t^* \mid \cdot)\right) \frac{q(\mathcal{I} \mid \boldsymbol{A}_{t+1}^*, \cdot)}{q(\mathcal{I} \mid \boldsymbol{A}_t^*, \cdot)}, 1\right).$$

5.2 Block Gibbs Algorithm

We propose a Block Gibbs approach that alternates between updating $\boldsymbol{A}^*$ with LIP and the continuous parameters $\boldsymbol{\Theta}$ using a continuous kernel $Q_c(\boldsymbol{\Theta} \mid \boldsymbol{A}^*, \boldsymbol{\mathcal{D}})$, such as HMC (Neal, 2011) or No-U-Turn-Sampler (NUTS, Hoffman et al., 2014). Algorithm 2 outlines the procedure. In each iteration, we first perform L updates of $\boldsymbol{A}^*$, for a maximum of $L \times K$ entry flips per iteration. Given its new network state $\boldsymbol{A}_{t+1}^*$, we perform a single update of the continuous parameters. The latter can also consist of multiple steps, for example, if an HMC with multiple integration steps is used (Neal, 2011). See Betancourt (2018) for more details regarding practical applications of HMC.

Algorithm 2 Block Gibbs Posterior Sampling

Input: Data $\boldsymbol{\mathcal{D}}$, Continuous kernel Q_c, total steps T, LIP steps L, entries per update K.

1: Initialize $\boldsymbol{A}_0^*, \boldsymbol{\Theta}_0$.
2: **for** $t = 0$ to $T - 1$ **do**
3: $\quad \boldsymbol{A}_{t+1}^* \leftarrow \boldsymbol{A}_t^*$.
4: $\quad$ **for** $\ell = 1$ to L **do** ▷ LIP updates using Algorithm 1
5: $\quad\quad \boldsymbol{A}_{t+1}^* \leftarrow \text{LIP}(\boldsymbol{A}_{t+1}^*, K \mid \boldsymbol{\Theta}_t, \boldsymbol{\mathcal{D}})$.
6: $\quad$ **end for**
7: $\quad \boldsymbol{\Theta}_{t+1} \leftarrow Q_c(\boldsymbol{\Theta}_t \mid \boldsymbol{A}_{t+1}^*, \boldsymbol{\mathcal{D}})$.
8: **end for**
9: **Return:** $\Big\{\boldsymbol{A}_t^*, \boldsymbol{\Theta}_t\Big\}_{t=1}^{T}$

There is an inherent trade-off in choosing L and K. Larger values allow for more extensive exploration of the network space per iteration, potentially improving mixing. However, they also increase the computational cost per iteration. Furthermore, performing excessive updates of $\boldsymbol{A}^*$ before updating $\boldsymbol{\Theta}$ may lead to drift in the Markov chains of the continuous

parameters, as changing the network structure can significantly alter the geometry of the conditional posterior of $\boldsymbol{\Theta}$.

We found that this trade-off can be mitigated by a robust initialization strategy. Specifically, we employ the following multi-stage procedure grounded in Bayesian modularization (Bayarri et al., 2009; Jacob et al., 2017) to ensure the chain starts in a high-probability region:

1. Estimate the parameters governing the network and proxy models, $(\boldsymbol{\theta}, \boldsymbol{\gamma})$, by sampling from the marginalized network module $p(\boldsymbol{\theta}, \boldsymbol{\gamma} \mid \mathcal{A}, \boldsymbol{X}) \propto p(\boldsymbol{\theta})p(\boldsymbol{\gamma}) \sum_{\boldsymbol{A}^*} p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma})p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta})$. This approach mirrors latent network reconstruction methods using proxies (Young et al., 2021). We denote the resulting estimates (e.g., posterior means) by $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$.

2. Conditionally on $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$, sample multiple networks from $p(\boldsymbol{A}^* \mid \mathcal{A}, \boldsymbol{X}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$ and select $\boldsymbol{A}_0^*$ as the network with the highest conditional probability.

3. Estimate the outcome model parameters $\boldsymbol{\eta}$ (and $\boldsymbol{\beta}$ if applicable) using $\boldsymbol{A}_0^*$ as a fixed covariate, yielding initial estimates $\boldsymbol{\Theta}_0$.

4. Finally, refine $\boldsymbol{A}_0^*$ by performing a short sequence of LIP updates (Algorithm 1) conditional on the initialized full continuous parameters $\boldsymbol{\Theta}_0$.

Steps 1–3 provide initial estimates using the cut-posterior distribution, which removes the 'feedback' between the network and outcome modules. The final LIP refinement step modifies $\boldsymbol{A}_0^*$ using information from the outcome and treatment models, effectively moving the initial state towards the higher density region of the full posterior. The resulting values are used to initialize Algorithm 2. Further technical details regarding this procedure are provided in Appendix C.

5.3 Causal Effects Estimation

We describe how to estimate the causal estimands (Section 2.3) given T posterior samples. For simplicity of presentation, we focus on the static policy. Following our Bayesian framework, let $\mathbb{E}\big[Y_i(\boldsymbol{z}) \mid \boldsymbol{\mathcal{D}}\big]$ be the expected outcome of unit i under the intervention that sets $\boldsymbol{Z} = \boldsymbol{z}$, given the observed data $\boldsymbol{\mathcal{D}}$. Under the SCMs described in Figures 1(a)–(c), we have

$$\begin{aligned}\mathbb{E}\big[Y_i(\boldsymbol{z}) \mid \boldsymbol{\mathcal{D}}\big] &= \mathbb{E}_{\boldsymbol{A}^*,\boldsymbol{\eta}}\Big[\mathbb{E}\big[Y_i(\boldsymbol{z}) \mid \boldsymbol{\mathcal{D}}, \boldsymbol{A}^*, \boldsymbol{\eta}\big] \mid \boldsymbol{\mathcal{D}}\Big] \\ &= \mathbb{E}_{\boldsymbol{A}^*,\boldsymbol{\eta}}\Big[\mathbb{E}\big[Y_i \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X}, \boldsymbol{A}^*, \boldsymbol{\eta}\big] \mid \boldsymbol{\mathcal{D}}\Big],\end{aligned}$$

where the second equality follows since all back-door paths between $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are blocked (Section 2). Under Figure 1(d) we have to condition on the proxies $\mathcal{A}$ as well, as described in Section 2.3. With a slight abuse of notation let $\widehat{\mu}_i(\boldsymbol{z} \mid \boldsymbol{\mathcal{D}})$ be the posterior expectation of $\mathbb{E}\big[Y_i(\boldsymbol{z}) \mid \boldsymbol{\mathcal{D}}\big]$. We can approximate this expectation using the posterior samples to obtain an estimator

$$\widehat{\mu}_i(\boldsymbol{z} \mid \boldsymbol{\mathcal{D}}) \approx T^{-1} \sum_{t=1}^{T} \mathbb{E}\big[Y_i \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X}, \boldsymbol{A}_t^*, \boldsymbol{\eta}_t\big], \tag{12}$$

where $\boldsymbol{A}_t^*, \boldsymbol{\eta}_t$, $t = 1, \ldots, T$ are the samples from the the posterior distribution $p(\boldsymbol{A}^*, \boldsymbol{\eta} \mid \boldsymbol{\mathcal{D}})$. If the inner conditional expectation $\mathbb{E}\big[Y_i \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X}, \boldsymbol{A}_t^*, \boldsymbol{\eta}_t\big]$ does not have a closed-form expression, we can approximate it via Monte Carlo by drawing Y_i from its posterior predictive distribution (Appendix C). Therefore, the point estimate of a static policy for unit i is $\widehat{\tau}_i(\boldsymbol{z}, \boldsymbol{z}' \mid \boldsymbol{\mathcal{D}}) = \widehat{\mu}_i(\boldsymbol{z} \mid \boldsymbol{\mathcal{D}}) - \widehat{\mu}_i(\boldsymbol{z}' \mid \boldsymbol{\mathcal{D}})$, and the population-level point estimate is $\widehat{\tau}(\boldsymbol{z}, \boldsymbol{z}' \mid \boldsymbol{\mathcal{D}}) = N^{-1} \sum_i \widehat{\tau}_i(\boldsymbol{z}, \boldsymbol{z}' \mid \boldsymbol{\mathcal{D}})$. In addition, credible intervals can be computed using the percentiles of

$$\widehat{\tau}^{(t)}(\boldsymbol{z}, \boldsymbol{z}' \mid \boldsymbol{\mathcal{D}}) = N^{-1} \sum_i \left(\mathbb{E}\big[Y_i \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X}, \boldsymbol{A}_t^*, \boldsymbol{\eta}_t\big] - \mathbb{E}\big[Y_i \mid \boldsymbol{Z} = \boldsymbol{z}', \boldsymbol{X}, \boldsymbol{A}_t^*, \boldsymbol{\eta}_t\big] \right),$$

over the $t = 1, \ldots, T$ posterior samples. Note that we can also write $\widehat{\tau}(\boldsymbol{z}, \boldsymbol{z}' \mid \boldsymbol{\mathcal{D}}) = T^{-1} \sum_{t=1}^{T} \widehat{\tau}^{(t)}(\boldsymbol{z}, \boldsymbol{z}' \mid \boldsymbol{\mathcal{D}})$, i.e., as the posterior mean. We can similarly obtain estimates and intervals for dynamic and stochastic estimands (Section 2.3).

6 Numerical Illustrations

We conducted numerical experiments using fully- and semi-synthetic data to evaluate the performance of our proposed methods. In the fully-synthetic scenario, all data components are simulated. In contrast, the semi-synthetic scenario combines real network data with simulated treatments and outcomes. We compare our proposed estimator based on the Block Gibbs sampler against several baselines. Key details and main results are presented below, while specific parameter settings and additional results are provided in Appendix D.

In both fully- and semi-synthetic scenarios, we assigned treatments independently to each unit with $\Pr(Z_i = 1 \mid \boldsymbol{A}^*) \propto d_i^*$, namely, treatment assignment was proportional to the degree of i in the true interference network. Outcomes were generated from a linear model with $\boldsymbol{Y} = \boldsymbol{\mu} + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$ are independent errors. The mean component $\boldsymbol{\mu}$ for unit i was $\mu_i = \eta_1 Z_i + \eta_2 \phi_1(\boldsymbol{Z}_{-i}, \boldsymbol{A}^*) + \boldsymbol{\eta}_x' \boldsymbol{X}_i$, and the exposure mapping $\phi_1(\mathbf{Z}_{-i}, \mathbf{A}^*)$ was set to summarize neighbors' treatments through a weighted average $\phi_1(\boldsymbol{Z}_{-i}, \boldsymbol{A}^*) = \sum_{j \neq i} w_j Z_j A_{ij}^*$. The weights were taken to be the degree centrality, $w_j = \frac{d_j^*}{N-1}$, assigning greater influence to units with higher network centrality. Here, η_1 represents the direct treatment effect, while η_2 captures the interference effect through neighbors' treatment.

6.1 Fully-Synthetic Data

We generated fully-synthetic data for $N = 500$ units. In each simulation iteration, we sampled data from the following SCM. Covariates $\boldsymbol{X}_i = (X_{1,i}, X_{2,i})$, were simulated from $X_{1,i} \sim N(0, 1)$ and $X_{2,i} \sim Ber(0.1)$. The edge-level covariates were defined as $\widetilde{X}_{1,ij} = \big|X_{1,i} - X_{1,j}\big|$, capturing the covariates distance, and $\widetilde{X}_{2,ij} = \mathbb{I}\{X_{2,i} + X_{2,j} = 1\}$, an indicator of exactly one of the two units having $X_2 = 1$. Edges in the latent interference network $\boldsymbol{A}^*$ were generated independently according to

$$\Pr(A_{ij}^* = 1 \mid \boldsymbol{X}, \boldsymbol{\theta}) = s\big(\theta_0 + \theta_1 \widetilde{X}_{1,ij} + \theta_2 \widetilde{X}_{2,ij}\big), \tag{13}$$

where $s(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. Parameters $\boldsymbol{\theta}$ were selected to produce sparse networks with a heavy-tailed degree distribution, reflecting common features of real-world

networks (see Appendix D for details). Given the true network $\boldsymbol{A}^*$, we generated two independent proxy networks $\mathcal{A} = \{\boldsymbol{A}_1, \boldsymbol{A}_2\}$ from a differential measurement error model

$$\Pr(A_{ij} = 1 \mid A^*_{ij}, \boldsymbol{X}, \boldsymbol{\gamma}) = s\big(A^*_{ij}\gamma_0 + (1 - A^*_{ij})(\gamma_1 + \gamma_2 \widetilde{X}_{1,ij} + \gamma_3 \widetilde{X}_{2,ij})\big). \tag{14}$$

We varied $\boldsymbol{\gamma}$ values to control the (dis)similarity between proxies and the true network. This model represents a scenario of causal proxies. Specifically, false-positive edges ($A_{ij} = 1, A^*_{ij} = 0$) occurred with probabilities that depend on covariate similarities, while true edges ($A^*_{ij} = 1$) were missing in $\boldsymbol{A}_1$ completely at random.

We took $N(0, 3^2)$ priors for all coefficients and $\sigma \sim \text{HalfNormal}(0, 3^2)$ for the noise term. In each iteration, we re-sampled all data from the SCM but fixed the true parameter values throughout. We compared two variants of the proposed Block Gibbs (BG) algorithm, using either a single ($\boldsymbol{A}_1$) or both proxies ($\mathcal{A}$), to several baselines. These include a naive estimation approach that takes the first proxy $\boldsymbol{A}_1$ ("Obs.") as the assumed interference network and the oracle that uses the true network $\boldsymbol{A}^*$ ("True"). Also, we tested the performance of continuous relaxation of $\boldsymbol{A}^*$ edges into $[0, 1]$ using the CONCRETE distribution (Maddison et al., 2017) ("Cont. relax"), which is a differentiable relaxation of the discrete Gumbel-Max trick. We also tested how using a two-stage approach that first estimates $\boldsymbol{A}^*$ from the proxies and then estimates the outcome model while treating the estimated network as fixed. In this approach, we further considered two methods to obtain the estimated fixed network: (i) using only the cut-posterior samples of $\boldsymbol{A}^*$ ("Two-stage (cut)"), corresponding to steps 1–3 of the BG initialization (Section 5.2), and (ii) refining the samples via several LIP steps ("Two-stage (LIP)"), which incorporates step 4 of the initialization strategy. We also assessed the contribution of the treatment signal to the inference by running a model without the treatment model ("BG (no Z)"). The impact of model misspecification was also evaluated by omitting the covariate $\boldsymbol{X}_1$ from all models ("BG (misspec)").

The fixed-network methods ("Obs.", "True", and two-stage methods) require only an outcome model and were estimated using NUTS (Hoffman et al., 2014). In the BG algorithms, we used NUTS for the continuous kernel, but also evaluated the performance of BG with HMC (Betancourt, 2018) for the continuous kernel ("BG (HMC)") when using a single proxy network. Each BG iteration included $L = 1$ LIP steps with $K = 1$ edge updates per step. Initial values were obtained from the cut-posterior followed by LIP refinement (Appendix C). The continuous relaxation method was estimated using Stochastic Variational Inference, see Appendix D for details. All MCMC methods used $T = 1.2 \times 10^4$ posterior samples ($3,000$ samples across four chains).

Our analysis evaluated dynamic and static policies (Section 2.3). For static policies, we considered the TTE, $\tau(\mathbf{1}, \mathbf{0})$, comparing the scenarios where all units are treated versus none. We considered the dynamic policy $h_{c,i}(X_1) = \mathbb{I}\{X_{1,i} \notin [-c, c]\}$, assigning treatment only to units whose covariate exceeds a threshold c. The dynamic estimand was taken to be $\tau(h_{c_1}, h_{c_2}) = \mu(h_{c_1}) - \mu(h_{c_2})$, with thresholds $c_1 = 0.75$ and $c_2 = 1.5$.

Let $\widehat{\tau}_i$ be the point estimate for units i of either the dynamic or static policies (Section 5.3). Estimation accuracy was assessed by unit-level Mean Absolute Percentage Error (MAPE) between posterior point estimates and true estimands τ_i, defined by $\text{MAPE} = N^{-1} \sum_{i \in [N]} \left| \frac{\hat{\tau}_i - \tau_i}{\tau_i} \right|$. Lower MAPE values indicate better estimation accuracy. We also evaluated the mean relative error of the population-level estimands, defined by $\mathbb{E}\left(\left| \frac{\hat{\tau} - \tau}{\tau} \right| \right)$. These

results are presented in Appendix D and exhibit performance patterns consistent with the unit-level MAPE.

Figure 3 summarizes the MAPE (mean $\pm$ standard deviation) values over 300 iterations across varying proxies strength (γ_2). As proxies become weaker (larger γ_2), naively estimating τ while treating the observed proxies as the true network leads to higher estimation error. In contrast, the BG algorithms consistently achieves accurate estimates. Specifically, the BG with either a single or two proxies attains the lowest MAPE across all proxy strengths and estimands compared to all baselines other than the true oracle. Using HMC ("BG (HMC)") instead of NUTS ("BG") for the continuous kernel in the BG sampler with one proxy network slightly increases MAPE. Ignoring the treatment model in the BG sampler ("BG (no Z)") slightly increases MAPE, illustrating the benefit of integrating all available data modules. Moreover, model misspecification ("BG (misspec)") only mildly affected the MAPE. The continuous relaxation approach also yielded high MAPE values. The two-stage methods performed worse than all of the BG samplers. However, the two-stage approaches with LIP refinement significantly outperformed the cut-posterior variant. Nevertheless, the two-stage methods (even with LIP refinement) tend to underestimate the uncertainty of the estimands, as reflected by empirical coverage rates of the 95% credible intervals (CIs) being below the nominal level (Appendix D). In contrast, the BG algorithms achieved empirical coverage rates close to the nominal level across all proxy strengths, demonstrating the necessity of the full iterative sampling procedure.

Figure 4 illustrates how the posterior distribution of $\boldsymbol{A}^*$ concentrates around the true network structure, by comparing the true network $\mathbf{A}^*$, the observed proxy network $\mathbf{A}_1$, and the posterior edge probabilities derived from the BG algorithm. The figure highlights that the posterior distribution effectively captures the true network topology even under weak proxy information ($\gamma_2 = 3$). Additional results, including traceplots, mixing diagnostics, and scaling BG to larger networks, are provided in Appendix D.

6.2 Semi-Synthetic Data

In the second experiment, we analyzed a multilayer network dataset collected by Magnani et al. (2013), which originally contained five layers describing social connections among $N = 61$ employees at Aarhus University's Department of Computer Science. Three layers are obtained through surveys asking employees about regular working relationships ("Work"), repeated leisure activities ("Leisure"), and eating lunch together ("Lunch"). Two additional layers captured online Facebook friendships and publication co-authorship. Since the co-authorship network is almost fully contained within other layers (Magnani et al., 2013), it was excluded from our analysis, leaving us with four unweighted and undirected networks. Summary statistics for each layer are presented in Table 1.

	Facebook	Leisure	Lunch	Work
Number of edges	124	88	193	194
Average degree	7.75	3.74	6.43	6.47
Number of isolated units	29	14	1	1

Table 1: Summary statistics of the multilayer network. Each column represents a layer.

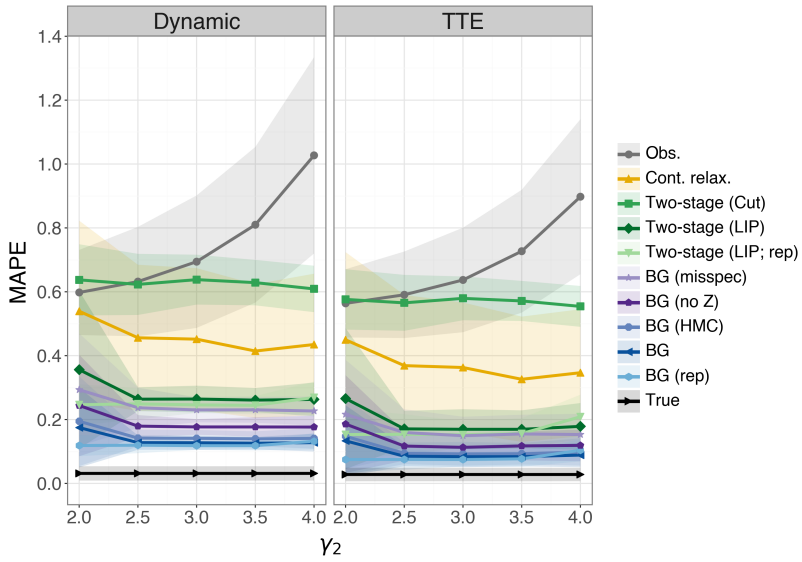


Figure 3: Mean ($\pm$ SD) of MAPE values across 300 simulation iterations for dynamic and TTE estimands. Proxy networks $\mathcal{A}$ are weaker as γ_2 (x-axis) increases. Smaller values indicate better performance.

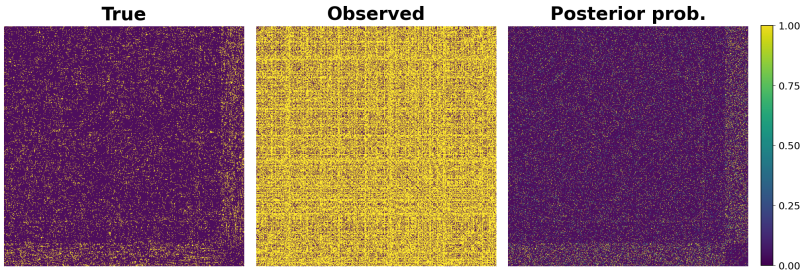


Figure 4: Heatmap of true $\boldsymbol{A}^*$, observed proxy $\boldsymbol{A}_1$, and posterior probabilities under one iteration with $\gamma_2 = 3$. In the True and Observed heatmaps, entries are binary. The posterior probabilities are the proportion of times an edge existed in the posterior network samples using a Block Gibbs sampler with a single proxy network. Nodes were first rearranged by hierarchical clustering for clearer visualization.

To illustrate the performance of our proposed method, we sequentially designated each layer as the latent interference network $\boldsymbol{A}^*$, treating the remaining three layers as observed proxies $\mathcal{A}$. The joint distribution of the four layers was modeled using LSM with shared bivariate latent positions $\boldsymbol{V}$

$$\Pr(A_{b,ij} = 1 \mid \boldsymbol{V}, \boldsymbol{\gamma}) = s\left(\gamma_{b,0} - e^{\gamma_{b,1}} \|\boldsymbol{V}_i - \boldsymbol{V}_j\|_2\right),$$

where $\gamma_{b,0}$ controls the baseline probability of edge creation in layer b and $\gamma_{b,1}$ controls how latent distances affects edge probabilities. We assumed independent Gaussian distribution for the latent positions $\boldsymbol{V}_i \sim N_2(\boldsymbol{0}, I)$, where I is the identity matrix, and hierarchical priors for the layer-specific parameters

$$\gamma_{b,j} \sim N(\mu_j, \sigma_j^2),\ \mu_j \sim N(0, 3^2),\ \sigma_j \sim \text{HalfNormal}(0, 3^2),\ j = 0, 1.$$

When a specific layer b is treated as the latent network $\boldsymbol{A}^*$, this model assumes that the three proxies indirectly inform its structure through the latent variables $\boldsymbol{V}$, μ_j, and σ_j. This setup represents non-causal proxies (Figure 1(c)), as none of the observed layers are assumed to directly influence the latent interference network and treatments depends only on $\boldsymbol{A}^*$. In each latent network scenario, treatments and outcomes were generated as in Section 6.1, although without covariates data. We focused on the TTE estimands, similar to Section 6.1. To produce a single "Observed" network from the three proxy layers, we aggregated edges using either the union ("OR") or intersection ("AND") operations. We compared the performance of the "True" network and the aggregated networks to the BG sampler using the same sampling setup as in Section 6.1. In addition, we included the two-stage method with LIP refinement as an additional baseline.

Table 2 summarizes the MAPE values across 300 iterations for each latent interference network layer and estimation method. The BG method consistently improved accuracy over other methods using the aggregated observed networks and two-stage methods with LIP refinement. Specifically, the BG method had estimation errors closer to those obtained with the true network. Two-stage methods with LIP refinement also improved over aggregation methods, but still lagged behind the BG sampler. Union aggregation ("OR") outperformed the intersection ("AND") in two of the four layers (Work and Lunch), while the opposite occurred in the other two layers (Facebook and Leisure). In addition, the BG method achieved empirical coverage rates close to the nominal level across all layers, while the Two-stage method and the aggregated networks did not (Appendix D).

7 Discussion

In this paper, we have introduced a novel framework for estimating causal effects when only proxy measurements of the latent interference network are available. Our proposed SCMs provide researchers flexibility in specifying models and choosing policy-relevant estimands. The latent interference network introduces considerable inferential challenges. Within a Bayesian framework, these challenges are translated to the posterior distribution being over a high-dimensional mixed parameter space composed of discrete (latent networks) and continuous (model parameters and latent variables) components. To address these challenges, we developed a Block Gibbs sampling algorithm that iteratively updates the continuous

Method	*Facebook*	*Leisure*	*Lunch*	*Work*
True	0.070 (0.045)	0.104 (0.080)	0.055 (0.041)	0.040 (0.024)
BG	0.132 (0.059)	0.157 (0.095)	0.111 (0.049)	0.096 (0.036)
Two-stage	0.397 (0.099)	0.187 (0.059)	0.169 (0.044)	0.287 (0.073)
OR	0.783 (0.198)	0.419 (0.229)	0.272 (0.057)	0.320 (0.065)
AND	0.501 (0.093)	0.233 (0.061)	0.425 (0.072)	0.324 (0.055)

Table 2: Mean (SD) of MAPE values across 300 iterations of the TTE estimand by layer and estimation method.

and discrete parameters. Specifically, discrete updates utilize Locally Informed Proposals, enabling efficient exploration of the large discrete latent network space. Crucially, these updates leverage all available data modules (e.g., proxy networks, outcomes, and treatments) to propose edge flips. This process actively integrates information from these diverse components, ensuring that the latent network is informed by all relevant data sources.

In this paper, we have classified the possible proxy networks into three types: measurement error, multiple sources, and multilayer networks (Section 3.1). This classification is not exhaustive but covers the proxy types compatible with our edge-level modeling framework, where each proxy network in $\mathcal{A}$ is an adjacency of the same dimension as the latent network $\boldsymbol{A}^*$. However, other forms of network-related proxies exist. For example, aggregated relational data, in which respondents report counts of their connections sharing specified traits rather than their exact identities (McCormick and Zheng, 2015), and proxies obtained from network sampling, in which only a subset of nodes and some of their edges are observed (Chandrasekhar and Lewis, 2011; Handcock and Gile, 2010). These designs yield aggregated information or partial data about the latent network rather than edge-level adjacency observations on the full unit set. Our framework conceptually accommodates such designs, e.g., by replacing the edge-level proxy model $p(\mathcal{A} \mid \boldsymbol{A}^*, \cdot)$ with a suitable aggregation likelihood. However, our inference algorithm (Section 5), and in particular the Locally Informed Proposals over the discrete edge set, is tailored to edge-level proxies and would require adaptations. We leave the integration of such proxy measurements to future work.

Our work can be extended to include estimands for peer effects and direct interventions on the network structure (Ogburn et al., 2024). Peer effects quantify how baseline outcomes, interpreted as "treatments", propagate through the interference network. Furthermore, estimands quantifying the effects of intervening on the interference network $\boldsymbol{A}^*$ structure can also be defined. However, even if $\boldsymbol{A}^*$ had been observed, in the presence of latent common causes $\boldsymbol{U}$ of $\boldsymbol{A}^*$ and $\boldsymbol{Y}$ (Figure 2), these estimands cannot generally be identified from observed data alone Instead, their estimation relies on additional modeling assumptions and prior specifications in our Bayesian framework.

As with all model-based methods, our approach relies on accurate model specification. The validity of these specifications can be assessed using prior and posterior predictive checks, which are integral to the Bayesian workflow (Gelman et al., 2020). We derive the posterior predictive distributions and describe how to draw from them in our settings in Appendix C. A promising direction for future research is extending our framework to gen-

eralized Bayesian methods (Bissiri et al., 2016; Matsubara et al., 2024), which can mitigate the impact of model misspecification on inference.

A natural question is whether applying our framework is detrimental when the observed proxy network is, in fact, an accurate measurement of the interference structure ($\boldsymbol{A} \approx \boldsymbol{A}^*$). As we show in an additional simulation study presented in Appendix D, when a single proxy coincides with the true network with high probability, the point estimates of the BG sampler are close to those of the oracle that conditions on $\boldsymbol{A}^*$, with the cost being variance inflation in the posterior. This mirrors a well-known phenomenon in the measurement error literature, where accounting for potential mismeasurement of accurately measured variables yields consistent but less efficient estimators (Carroll et al., 2006). In our case, because the BG sampler does not condition on the true network, it must additionally infer $\boldsymbol{\gamma}, \boldsymbol{\theta}$, and the latent network $\boldsymbol{A}^*$ itself, thus propagating the posterior uncertainty about $\boldsymbol{A}^*$ into the causal estimates. Since practitioners rarely know a priori whether $\boldsymbol{A} \approx \boldsymbol{A}^*$, this variance inflation is the price of robustness across the full spectrum of proxy quality, while avoiding the substantial bias that can arise from naively treating a noisy proxy as the truth.

While our method performs well for small to moderate network sizes, the dimension of the latent network space grows quadratically $\mathcal{O}(N^2)$ with the number of units N. Consequently, the computational burden of recovering the full posterior distribution via MCMC increases significantly for large networks. Adapting our framework to large-scale settings remains a challenging area for future research. Potential directions include developing more scalable approximation algorithms or shifting towards estimators that are robust to network misspecification without requiring full latent network recovery (e.g., by adapting the network-misspecification-robust design-based estimator of Weinstein and Nevo, 2026, to model-based settings).

Our findings also carry broader implications for recent methodological advancements. Existing techniques that rely on flexible outcome modeling, such as Graph Neural Networks (Ma and Tresp, 2021), or those proposing novel experimental designs for networked experiments (Ugander et al., 2013; Eckles et al., 2017) typically assume that the true interference network is observed. Our analysis highlights the practical reality that researchers often have only proxies. Extending these methods presents a valuable avenue for future research.

Acknowledgments and Disclosure of Funding

The authors gratefully acknowledge support from the Israel Science Foundation (ISF grant No. 2300/25). BW is supported by the Data Science Fellowship granted by the Israeli Council for Higher Education. We also thank three anonymous reviewers for helpful comments and suggestions that improved the paper.

Appendix – Table of Contents

Appendix A. Structural Causal Models

We describe here the SCMs corresponding to the other DAGs in Figure 2, extending the description of the SCMs of Figure 1(a) given in (1). Throughout this section, we follow the main text and assume that $f_U, f_X, f_{A^*}, f_{A_b}, f_Z, f_Y$ are fixed functions and $\boldsymbol{\varepsilon}_U, \boldsymbol{\varepsilon}_X, \boldsymbol{\varepsilon}_{A^*}, \boldsymbol{\varepsilon}_{A_b}, \boldsymbol{\varepsilon}_Z, \boldsymbol{\varepsilon}_Y$ are noise terms. We assume the pairwise noise vectors are independent, e.g., $\boldsymbol{\varepsilon}_Z \perp\!\!\!\perp \boldsymbol{\varepsilon}_Y$, but do not limit the dependence structure between units (e.g., the dependence of ε_{Y_i} and ε_{Y_j}). Furthermore, we write the structural equations for proxies $\boldsymbol{A}_b$ separately for each proxy, as in (1), but extending this for other scenarios, e.g., auto-regressive proxies, is possible as described in Section 2.2. Also, each of the outcome models f_Y can be modified to depend on summarizing functions ϕ_1, ϕ_2, ϕ_3 as in (2). Note that the causal estimands described in Section 2.3 are interventions on $\boldsymbol{Z}$ and therefore their interpretation remains the same regardless of whether the proxies are causal or not and whether treatment assignment is based on the latent network or the proxies.

Furthermore, we describe SCMs where the treatment assignment f_Z is influenced by both the true latent interference network $\boldsymbol{A}^*$ and the observed proxies $\mathcal{A}$. We consider this setting under scenarios of causal proxies ($\boldsymbol{A}^* \to \mathcal{A}$) and non-causal proxies ($\boldsymbol{A}^* \leftarrow \boldsymbol{V} \to \mathcal{A}$).

Causal proxies and proxy-based treatment assignment. The structural equations corresponding to Figure 1(b) are

$$
\begin{aligned}
\boldsymbol{U}_i &= f_U\big(\boldsymbol{\varepsilon}_{U_i}\big), && i \in [N] \\
\boldsymbol{X}_i &= f_X\big(\boldsymbol{\varepsilon}_{X_i}\big), && i \in [N] \\
\boldsymbol{A}^* &= f_{A^*}\big(\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{\varepsilon}_{A^*}\big), && \\
\boldsymbol{A}_b &= f_{A_b}\big(\boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\varepsilon}_{A_b}\big), && b \in [B] \\
Z_i &= f_Z\big(\mathcal{A}, \boldsymbol{X}_i, \boldsymbol{X}_{-i}, \varepsilon_{Z_i}\big), && i \in [N] \\
Y_i &= f_Y\big(Z_i, \boldsymbol{X}_i, \boldsymbol{Z}_{-i}, \boldsymbol{X}_{-i}, \boldsymbol{A}^*, \boldsymbol{U}_i, \varepsilon_{Y_i}\big) && i \in [N].
\end{aligned}
$$

Non-causal proxies and latent-based treatment assignment. The structural equations corresponding to Figure 1(c) are

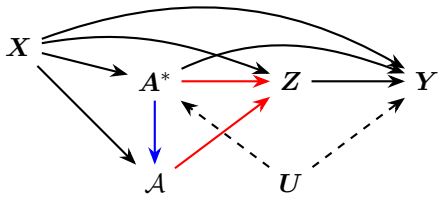
$$
\begin{aligned}
\boldsymbol{U}_i &= f_U\big(\boldsymbol{\varepsilon}_{U_i}\big), && i \in [N] \\
\boldsymbol{X}_i &= f_X\big(\boldsymbol{\varepsilon}_{X_i}\big), && i \in [N] \\
\boldsymbol{V}_i &= f_V\big(\boldsymbol{\varepsilon}_{V_i}\big), \\
\boldsymbol{A}^* &= f_{A^*}\big(\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{\varepsilon}_{A^*}\big), \\
\boldsymbol{A}_b &= f_{A_b}\big(\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{\varepsilon}_{A_b}\big), && b \in [B] \\
Z_i &= f_Z\big(\boldsymbol{A}^*, \boldsymbol{X}_i, \boldsymbol{X}_{-i}, \varepsilon_{Z_i}\big), && i \in [N] \\
Y_i &= f_Y\big(Z_i, \boldsymbol{X}_i, \boldsymbol{Z}_{-i}, \boldsymbol{X}_{-i}, \boldsymbol{A}^*, \boldsymbol{U}_i, \varepsilon_{Y_i}\big) && i \in [N],
\end{aligned}
$$

where f_V is a fixed function and $\boldsymbol{\varepsilon}_{V_i}$ are noise terms similar to other mentioned above. Each of $\boldsymbol{V}_i$ can be a vector of (latent) variables.

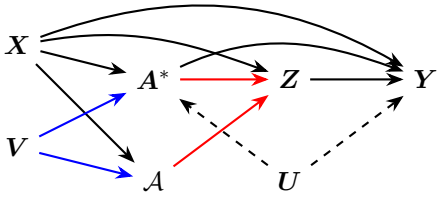
Non-causal proxies and proxy-based treatment assignment. The structural equations corresponding to Figure 1(d) are

$$
\begin{aligned}
\boldsymbol{U}_i &= f_U\big(\boldsymbol{\varepsilon}_{U_i}\big), && i \in [N] \\
\boldsymbol{X}_i &= f_X\big(\boldsymbol{\varepsilon}_{X_i}\big), && i \in [N] \\
\boldsymbol{V}_i &= f_V\big(\boldsymbol{\varepsilon}_{V_i}\big), \\
\boldsymbol{A}^* &= f_{A^*}\big(\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{\varepsilon}_{A^*}\big), \\
\boldsymbol{A}_b &= f_{A_b}\big(\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{\varepsilon}_{A_b}\big), && b \in [B] \\
Z_i &= f_Z\big(\mathcal{A}, \boldsymbol{X}_i, \boldsymbol{X}_{-i}, \varepsilon_{Z_i}\big), && i \in [N] \\
Y_i &= f_Y\big(Z_i, \boldsymbol{X}_i, \boldsymbol{Z}_{-i}, \boldsymbol{X}_{-i}, \boldsymbol{A}^*, \boldsymbol{U}_i, \varepsilon_{Y_i}\big) && i \in [N].
\end{aligned}
$$

Combination of proxy-based and latent-based treatment assignment. In the main text, we distinguished between treatment assignment based on the true latent network $\boldsymbol{A}^*$ and that based on the observed proxies $\mathcal{A}$. However, combinations of both are possible. The prominent scenarios are observational studies where researchers do not control the treatment assignment mechanism. For example, when the true latent network represents physical interactions while the observed proxies are networks of online interactions, treatment assignment can be influenced by both, while only the physical interactions (true network) correspond to the interference network through which treatments propagate to the outcomes. That is, treatments can be influenced by both online and physical interactions, e.g., by depending on both the online and physical nodal degrees. In that case, f_Z is influenced by both $\boldsymbol{A}^*$ and $\mathcal{A}$, but the outcomes are influenced only by $\boldsymbol{A}^*$. We describe the two DAGs corresponding to these scenarios with causal and non-causal proxy networks and show that this modification does not change the conditioning sets required to block the back-door paths between $\boldsymbol{Z}$ and $\boldsymbol{Y}$.



(a) Causal proxies & Combined assignment



(b) Non-causal proxies & Combined assignment

Figure A.1: DAGs representing the assumed structural models for two additional settings, corresponding to causal or non-causal proxies and combined treatment assignment based on both the latent network $\boldsymbol{A}^*$ and the proxies $\mathcal{A}$: (a) Causal proxies. (b) Non-causal proxies. The dashed arrows are optional.

Figure A.1 shows the DAGs in the two scenarios where treatment assignment is influenced by both $\boldsymbol{A}^*$ and $\mathcal{A}$. In Figure A.1(a)), the proxies $\mathcal{A}$ are direct descendants of the true latent network $\boldsymbol{A}^*$, while in Figure A.1(b)), both $\boldsymbol{A}^*$ and $\mathcal{A}$ are descendants of common latent variables $\boldsymbol{V}$ and do not share a direct connection.

In both cases, the only change to the structural equations is the treatment assignment model, which now has the form of

$$Z_i = f_Z\big(\mathcal{A}, \boldsymbol{A}^*, \boldsymbol{X}_i, \boldsymbol{X}_{-i}, \varepsilon_{Z_i}\big), \quad i \in [N].$$

Similar to the DAGs presented in the main text, the conditioning set required to block all back-door paths between $\boldsymbol{Z}$ and $\boldsymbol{Y}$ is $\boldsymbol{A}^*$ and $\boldsymbol{X}$ in Figure A.1(a)). In addition, similarly to Figure 1(d)) where $\boldsymbol{A}^*$ is a collider on the path $\boldsymbol{Y} \leftarrow \boldsymbol{U} \rightarrow \boldsymbol{A}^* \leftarrow \boldsymbol{V} \rightarrow \mathcal{A} \rightarrow \boldsymbol{Z}$, the conditioning set in Figure A.1(b)) is $\boldsymbol{X}, \boldsymbol{A}^*$, and $\mathcal{A}$.

Therefore, the only practical difference of this combined assignment setting to the those described in the main text is that the probabilistic model for $\boldsymbol{Z}$ should include both $\boldsymbol{A}^*$ and $\mathcal{A}$.

Appendix B. Identifiability of Structural Parameters

We consider the problem of identifying the structural parameters $\boldsymbol{\Theta} = (\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\theta})$ from the observed data $\boldsymbol{\mathcal{D}} = (\boldsymbol{Y}, \boldsymbol{Z}, \mathcal{A}, \boldsymbol{X})$. Let $\mathbf{m}(\boldsymbol{\mathcal{D}})$ be a vector of observable moments (e.g., means and cross-covariances). Denote the parameter space as Ω_Θ. We assume it is a subset of $\mathbb{R}^p$ for some $p > 0$. We define the *moment map* $\mathcal{M} : \Omega_\Theta \rightarrow \mathbb{R}^k$ as the expected value of these moments implied by the model parameters:

$$\mathcal{M}(\boldsymbol{\Theta}) = \mathbb{E}_{\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}}[\mathbf{m}(\boldsymbol{\mathcal{D}})]. \tag{B.1}$$

Global identification of $\boldsymbol{\Theta}$ can be established by showing that the moment map $\mathcal{M}(\boldsymbol{\Theta})$ is injective over the parameter space Ω_Θ (Rothenberg, 1971).

Definition B.1 (Global Identification via Moments)**.** *The parameter $\boldsymbol{\Theta}$ is globally identified with respect to the moments $\mathbf{m}(\boldsymbol{\mathcal{D}})$ if the map $\mathcal{M}(\boldsymbol{\Theta})$ is injective on the parameter space*

Ω_Θ. *That is, for any* $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \Omega_\Theta$*:*

$$\mathcal{M}(\boldsymbol{\Theta}_1) = \mathcal{M}(\boldsymbol{\Theta}_2) \implies \boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2.$$

Establishing injectivity of $\mathcal{M}(\boldsymbol{\Theta})$ can be challenging in general. However, we can leverage the structure of the SCMs and the conditional independencies it implies to derive explicit moment conditions that relate the observable moments to the parameters $\boldsymbol{\Theta}$.

Each of the observed variables $\boldsymbol{Y}$, $\boldsymbol{Z}$, and $\mathcal{A}$ can provide distinct information about the structural parameters $\boldsymbol{\Theta}$. The conditional independencies between these variables given $\boldsymbol{A}^*$ and $\boldsymbol{X}$, imply that any dependence observed in the data must be mediated through $\boldsymbol{A}^*$. Therefore, by analyzing the covariances and higher-order moments involving these variables, we can triangulate the information about the parameters $\boldsymbol{\Theta}$. This moment-based perspective aligns with the identification results for latent variable models using three independent views (Allman et al., 2009).

For example, consider the settings of causal proxies with a latent-based treatment assignment (Figure 1(a)), which implies that $\boldsymbol{Y} \perp\!\!\!\perp \mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}$ and $\boldsymbol{Z} \perp\!\!\!\perp \mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}$. Let $\boldsymbol{\mu}_{\boldsymbol{Y}}(\boldsymbol{A}^*; \boldsymbol{\eta}) = \mathbb{E}_{\boldsymbol{\eta}}[\boldsymbol{Y} \mid \boldsymbol{A}^*, \boldsymbol{X}]$, $\boldsymbol{\mu}_{\boldsymbol{Z}}(\boldsymbol{A}^*; \boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{Z} \mid \boldsymbol{A}^*, \boldsymbol{X}]$, and $\boldsymbol{\mu}_{\mathcal{A}}(\boldsymbol{A}^*; \boldsymbol{\gamma}) = \mathbb{E}_{\boldsymbol{\gamma}}[\text{vec}(\mathcal{A}) \mid \boldsymbol{A}^*, \boldsymbol{X}]$ be the conditional means of $\boldsymbol{Y}$, $\boldsymbol{Z}$, and $\mathcal{A}$ given $\boldsymbol{A}^*$ and $\boldsymbol{X}$, where $\text{vec}(\mathcal{A})$ denotes the vectorization of the observed proxy networks $\mathcal{A}$, and when there are multiple proxies, we compute each vectorization separately. By the law of total covariance and the conditional independence statements above, we have the following system of covariance moments relating the observable moments to the parameters $\boldsymbol{\Theta}$:

$$\begin{aligned} \text{Cov}(\boldsymbol{Y}, \text{vec}(\mathcal{A}) \mid \boldsymbol{X}) &= \text{Cov}_{\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta}} \left(\boldsymbol{\mu}_{\boldsymbol{Y}}(\boldsymbol{A}^*; \boldsymbol{\eta}), \boldsymbol{\mu}_{\mathcal{A}}(\boldsymbol{A}^*; \boldsymbol{\gamma}) \right) \\ \text{Cov}(\boldsymbol{Z}, \text{vec}(\mathcal{A}) \mid \boldsymbol{X}) &= \text{Cov}_{\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta}} \left(\boldsymbol{\mu}_{\boldsymbol{Z}}(\boldsymbol{A}^*; \boldsymbol{\beta}), \boldsymbol{\mu}_{\mathcal{A}}(\boldsymbol{A}^*; \boldsymbol{\gamma}) \right) \end{aligned} \tag{B.2}$$

In addition to the above moments, we can also consider higher-order moments or other cross-moments. For instance, $\mathbb{E}[Y_i Y_j \mid \boldsymbol{X}]$, $\mathbb{E}[Y_i Y_j \mid A_{ij} = k, \boldsymbol{X}]$ for $k \in \{0, 1\}$, $\text{Cov}(\text{vec}(\boldsymbol{A}_1), \text{vec}(\boldsymbol{A}_2) \mid \boldsymbol{X})$ when $\mathcal{A}$ contains at least two proxies, and $\mathbb{E}[Z_i Z_j \mid \boldsymbol{X}]$. These additional moments can provide further constraints on the parameters $\boldsymbol{\Theta}$, potentially aiding in establishing global identification. The specific choice of moments will depend on the model structure and the available data.

We now turn to several examples of simplified models where we can explicitly derive the moment conditions and establish global identification. Specifically, we consider a class of models where the true network $\boldsymbol{A}^*$ is generated from a homogeneous Erdős-Rényi model, the observed proxy network $\mathcal{A}$ is a noisy measurement of $\boldsymbol{A}^*$, and the outcomes $\boldsymbol{Y}$ follows a linear model where exposures ϕ_1 is the number of treated neighbors.

B.1 Examples of Global Identification

Assume the following model:

$$\begin{aligned} A^*_{ij} &\sim \text{Ber}(\theta) \\ A_{ij} \mid A^*_{ij} = k &\sim \text{Ber}(\gamma_k), \quad k = 0, 1 \\ Z_i &\sim \text{Ber}(p_z) \\ Y_i &= \eta_1 Z_i + \eta_2 \sum_{j \neq i} A^*_{ij} Z_j + \varepsilon_i, \end{aligned} \tag{B.3}$$

where ε_i are independent and zero mean error terms with $\mathbb{E}[\varepsilon_i \mid \boldsymbol{Z}, \boldsymbol{A}^*] = 0$. In this model, the structural parameters are $\boldsymbol{\Theta} = (\theta, \gamma_0, \gamma_1, p_z, \eta_1, \eta_2)$. The true network $\boldsymbol{A}^*$ is latent. We seek to recover the parameters $\boldsymbol{\Theta}$ from the observed data $\boldsymbol{\mathcal{D}} = (\boldsymbol{Y}, \boldsymbol{Z}, \mathcal{A})$. If we are able to do so, then we can recover the true network model $p(\boldsymbol{A}^* \mid \theta)$, and therefore the distribution of causal estimands (see Remark 1). The simple model (B.3) is an example of causal proxies (Figure 1(a)-(b)), when there is one proxy and no covariates $\boldsymbol{X}$ or latent variables $\boldsymbol{V}$.

We consider here different extensions of the model (B.3) and investigate their identifiability. To establish global identification, we utilize Definition B.1 and determine whether the system of moment equations has a unique solution, thus ensuring that the moment map $\mathcal{M}(\boldsymbol{\Theta})$ is injective.

Starting with p_z, obviously, we can write $\mathbb{E}[Z_i] = p_z$ with the marginal first moment of Z_i. Also, note that

$$\mathbb{E}[Y_i \mid Z_i = z] = \eta_1 z + \eta_2 (N-1) p_z \theta, \tag{B.4}$$

therefore,

$$\mathbb{E}[Y_i] = \eta_1 p_z + \eta_2 (N-1) p_z \theta, \tag{B.5}$$

and we can represent η_1 by

$$\eta_1 = \mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]. \tag{B.6}$$

We also have

$$\mu_A = \Pr(A_{ij} = 1) = \gamma_1 \theta + \gamma_0 (1 - \theta). \tag{B.7}$$

Let $d_i^{obs} = \sum_{j \neq i} A_{ij}$ be the observed degree of unit i in the proxy network. Look at covariance between Y_i and d_i^{obs},

$$\begin{aligned}
\operatorname{Cov}(Y_i, d_i^{obs}) &= \operatorname{Cov}(\eta_1 Z_i + \eta_2 \sum_{j \neq i} A_{ij}^* Z_j, \sum_{j \neq i} A_{ij}) \\
&= \eta_2 \sum_{j \neq i} \operatorname{Cov}(A_{ij}^* Z_j, A_{ij}) \\
&= \eta_2 (N-1) p_z \theta (\gamma_1 - \mu_A),
\end{aligned}$$

where we used the fact that $\mathbb{E}[Z_j A_{ij} A_{ij}^*] = p_z \theta \gamma_1$ and $\mathbb{E}[Z_j A_{ij}^*] = p_z \theta$. Therefore, we can write η_2 as

$$\eta_2 = \frac{\operatorname{Cov}(Y_i, d_i^{obs})}{(N-1) p_z \theta (\gamma_1 - \mu_A)} = \frac{\operatorname{Cov}(Y_i, d_i^{obs})}{(N-1) p_z \theta (1-\theta)(\gamma_1 - \gamma_0)}. \tag{B.8}$$

Omitting p_z and η_1, as they are trivially identified, we focus on the remaining parameters $(\theta, \gamma_0, \gamma_1, \eta_2)$. However, we have three equations (B.5), (B.7), and (B.8) for four unknowns, so without further assumptions the model (B.3) is not identified. Consider the following extension to (B.3) that yield global identification.

Known γ_0 or $\gamma_0 = f(\gamma_1)$. If either γ_0 is known or there exists a known function f such that $\gamma_0 = f(\gamma_1)$, then we can solve the three equations for the three unknowns $(\theta, \gamma_1, \eta_2)$. For example, if we know that $\gamma_0 = 0$, then from (B.7) we have $\gamma_1 = \mu_A / \theta$, and substituting this into (B.8) gives $\eta_2 = \operatorname{Cov}(Y_i, d_i^{obs}) / \big((N-1) p_z \mu_A (1-\theta)\big)$. We can then substitute η_2 into (B.4) to solve for θ.

Two proxy networks. Assume we have two proxy networks $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ independently generated given $\boldsymbol{A}^*$ as in (B.3) with the same parameters γ_0 and γ_1. Thus, given $A^*_{ij} = 1$ we have

$$|A_{1,ij} - A_{2,ij}| = \begin{cases} -1 & \text{w.p. } \gamma_1(1-\gamma_1) \\ 0 & \text{w.p. } 1 - 2\gamma_1(1-\gamma_1) \\ 1 & \text{w.p. } \gamma_1(1-\gamma_1), \end{cases}$$

and given $A^*_{ij} = 0$ we have

$$|A_{1,ij} - A_{2,ij}| = \begin{cases} -1 & \text{w.p. } \gamma_0(1-\gamma_0) \\ 0 & \text{w.p. } 1 - 2\gamma_0(1-\gamma_0) \\ 1 & \text{w.p. } \gamma_0(1-\gamma_0). \end{cases}$$

Therefore, marginally we have

$$\Pr\left(|A_{1,ij} - A_{2,ij}| = 1\right) = 2\left[\gamma_1(1-\gamma_1)\theta + \gamma_0(1-\gamma_0)(1-\theta)\right].$$

Define the mean absolute difference between the two proxy networks as

$$m_{AD} = \frac{2}{N(N-1)} \sum_{i<j} |A_{1,ij} - A_{2,ij}|.$$

Then, $\mathbb{E}[m_{AD}] = 2\left[\gamma_1(1-\gamma_1)\theta + \gamma_0(1-\gamma_0)(1-\theta)\right]$. Therefore, we have four equations (B.5), (B.7), (B.8), and the latter expectation for m_{AD} for four unknowns $(\theta, \gamma_0, \gamma_1, \eta_2)$, and the model is globally identified.

Treatment as a function of $\boldsymbol{A}^*$. Assume that treatment assignment depends on the true network $\boldsymbol{A}^*$. Specifically, assume that independently for each unit i,

$$\Pr(Z_i = 1 \mid \boldsymbol{A}^*) = p_z \frac{d^*_i}{\mathbb{E}[d^*_i]} = p_z \frac{d^*_i}{(N-1)\theta} \propto d^*_i,$$

where $d^*_i = \sum_{j \neq i} A^*_{ij}$ is the degree of unit i in the true network. We can identify p_z from the marginal first moment $\mathbb{E}[Z_i] = p_z$. Note that Z_i and Z_j are independent only given $\boldsymbol{A}^*$, but marginally they are dependent. Let $\kappa = \frac{p_z}{(N-1)\theta}$. We have

$$\begin{aligned}
\mathbb{E}[Z_i Z_j] &= \mathbb{E}\left[\mathbb{E}[Z_i Z_j \mid \boldsymbol{A}^*]\right] \\
&= \mathbb{E}\left[\left(\kappa \sum_{k \neq i} A^*_{ik}\right)\left(\kappa \sum_{k \neq j} A^*_{jk}\right)\right] \\
&= \kappa^2 \mathbb{E}\left[(A^*_{ij})^2 + ((N-1)^2 - 1) A^*_{ik} A^*_{jl}\right] \\
&= \kappa^2 \left[\theta + ((N-1)^2 - 1)\theta^2\right] \\
&= \frac{p_z^2}{(N-1)^2\theta^2}\left[\theta + ((N-1)^2 - 1)\theta^2\right] \\
&= \frac{p_z^2}{(N-1)^2}\left[\frac{1}{\theta} + ((N-1)^2 - 1)\right]
\end{aligned}$$

solving for θ gives

$$\theta = \frac{p_z^2}{(N-1)^2(\mathbb{E}[Z_i Z_j] - p_z^2) + p_z^2}.$$

With this expression, we can solve for η_2 using (B.5). Then, solve for γ_1 and γ_0 using (B.7) and (B.8). Furthermore, given $\boldsymbol{A}^*$, the treatment Z_i is independent of the observed degree d_i^{obs}, therefore, using the law of total covariance, we can express the covariance between Z_i and d_i^{obs} as

$$\begin{aligned}
Cov(Z_i, d_i^{obs}) &= Cov\left(\mathbb{E}[Z_i \mid \boldsymbol{A}^*], \mathbb{E}[d_i^{obs} \mid \boldsymbol{A}^*]\right) \\
&= Cov\left(\kappa \sum_{j \neq i} A_{ij}^*, \sum_{j \neq i} [A_{ij}^* \gamma_1 + (1 - A_{ij}^*)\gamma_0]\right) \\
&= Cov\left(\kappa \sum_{j \neq i} A_{ij}^*, \sum_{j \neq i} A_{ij}^* (\gamma_1 - \gamma_0)\right) \\
&= (N-1)\kappa\theta(1-\theta)(\gamma_1 - \gamma_0) \\
&= p_z(1-\theta)(\gamma_1 - \gamma_0).
\end{aligned}$$

Thus, we can also express $\gamma_1 - \gamma_0 = \frac{Cov(Z_i, d_i^{obs})}{p_z(1-\theta)}$, then solve for η_2 using (B.8) and finally solve for γ_0 and γ_1 using (B.7). This implies that the structural parameters are over-identified in this scenario.

Network-correlated errors. Assume that the error terms ε_i are still mean zero but have a covariance structure $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \rho A_{ij}^*$ for some $\rho > 0$. These represent a neighborhood correlation structure in the outcomes. The parameters (other than p_z and η_1) are now $(\theta, \gamma_0, \gamma_1, \eta_2, \rho)$. We have,

$$Cov(Y_i, Y_j) = \eta_2^2 (N-2)\theta^2 p_z(1-p_z) + \rho\theta. \tag{B.9}$$

Unlike the independent error case, this moment depends on ρ. Next, we define the auxiliary variable $R_{ij} = \eta_1 Z_i + \eta_2 \sum_{k \neq i,j} A_{ik}^* Z_k$. We can write $Y_i = R_{ij} + \eta_2 A_{ij}^* Z_j + \varepsilon_i$. Thus, $\mathbb{E}[R_{ij}] = \mathbb{E}[Y_i] - \eta_2 \theta p_z$. Also, note that $\Pr(A_{ij}^* = 1 \mid A_{ij} = 1) = \frac{\theta\gamma_1}{\mu_A}$ and $\Pr(A_{ij}^* = 1 \mid A_{ij} = 0) = \frac{\theta(1-\gamma_1)}{1-\mu_A}$. Then, we have for $k = 0, 1$

$$\begin{aligned}
\mathbb{E}[Y_i Y_j \mid A_{ij} = k] &= \mathbb{E}\left[(R_{ij} + \eta_2 A_{ij}^* Z_j + \varepsilon_i)(R_{ji} + \eta_2 A_{ij}^* Z_i + \varepsilon_j) \mid A_{ij} = k\right] \\
&= \mathbb{E}[R_{ij} R_{ji}] + \mathbb{E}\left[\mathbb{E}\left[2R_{ij}\eta_2 A_{ij}^* Z_j + (\eta_2 A_{ij}^*)^2 Z_i Z_j + \varepsilon_i \varepsilon_j \mid A_{ij}^*\right] \mid A_{ij} = k\right] \\
&= \mathbb{E}[R_{ij} R_{ji}] + \Pr(A_{ij}^* = 1 \mid A_{ij} = k)\left[2\eta_2 p_z(\mathbb{E}[Y_i] - \eta_2 \theta p_z) + \eta_2^2 p_z^2 + \rho.\right]
\end{aligned}$$

Taking the difference for $k = 1$ and $k = 0$ gives

$$\mathbb{E}[Y_i Y_j \mid A_{ij} = 1] - \mathbb{E}[Y_i Y_j \mid A_{ij} = 0] = \left[\frac{\gamma_1 \theta}{\mu_A} - \frac{(1-\gamma_1)\theta}{1-\mu_A}\right]\left[2\eta_2 p_z(\mathbb{E}[Y_i] - \eta_2 \theta p_z) + \eta_2^2 p_z^2 + \rho.\right] \tag{B.10}$$

This provides five equations for the five unknowns $(\theta, \gamma_0, \gamma_1, \eta_2, \rho)$, thereby achieving global identification. The system formed by equations (B.5), (B.7), (B.8), (B.9), and (B.10) provides five independent equations for the five unknowns $(\theta, \gamma_0, \gamma_1, \eta_2, \rho)$, thereby achieving global identification.

B.2 Local Identifiability

We now discuss local identifiability of the model parameters $\boldsymbol{\Theta}$. Local identifiability is weaker than global identifiability. It means that in a neighborhood of the true parameter value $\boldsymbol{\Theta}_0$, there are no other observationally equivalent parameter values (Rothenberg, 1971). That is, there is identifiability only in a small region around $\boldsymbol{\Theta}_0$. A standard approach to establish local identifiability is through the Fisher Information Matrix (FIM). Specifically, Rothenberg (1971, Theorem 1) established that $\boldsymbol{\Theta}$ is locally identifiable at $\boldsymbol{\Theta}_0$ if and only if the FIM is nonsingular at $\boldsymbol{\Theta}_0$. We describe here the general conditions for local identifiability in our setting with latent network $\boldsymbol{A}^*$. We then provide some intuition on when these conditions hold. This result is formal and complements the global identification arguments presented earlier. However, verifying local identifiability through the FIM in practice is challenging due to the intractability of the observed data likelihood.

Consider for simplicity the causal proxies and latent-based treatment assignment (Figure 1(a)). Let $\boldsymbol{\Theta} = (\boldsymbol{\eta}, \boldsymbol{\beta}_Z, \boldsymbol{\gamma}, \boldsymbol{\theta})$ be the vector of all model parameters. Omitting covariates for simplicity, the observed data are $\boldsymbol{\mathcal{D}} = (\boldsymbol{Y}, \boldsymbol{Z}, \mathcal{A})$. The latent network $\boldsymbol{A}^*$ is missing. In this case, the complete data (if $\boldsymbol{A}^*$ were observed) likelihood factorizes as

$$p(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta}) = p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{\eta}) p(\boldsymbol{Z} \mid \boldsymbol{A}^*, \boldsymbol{\beta}_Z) p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{\gamma}) p(\boldsymbol{A}^* \mid \boldsymbol{\theta}). \tag{B.11}$$

However, the observed data likelihood is $p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\Theta}) = \sum_{\boldsymbol{A}^*} p(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta})$, which is impossible to compute in closed form for most models of interest. The complete data score function is defined as

$$\boldsymbol{S}_{comp}(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta}) = \nabla_{\boldsymbol{\Theta}} \log p(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta}). \tag{B.12}$$

Similarly, the observed data score function is

$$\boldsymbol{S}_{obs}(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\Theta}) = \nabla_{\boldsymbol{\Theta}} \log p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\Theta}). \tag{B.13}$$

We can represent the observed data score function (B.13) as the posterior expected complete data score function (B.12). Since $\nabla_{\boldsymbol{\Theta}} p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\Theta}) = \sum_{\boldsymbol{A}^*} \nabla_{\boldsymbol{\Theta}} p(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta})$, we have

$$\begin{aligned}
\boldsymbol{S}_{obs}(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\Theta}) &= \frac{\nabla_{\boldsymbol{\Theta}} p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\Theta})}{p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\Theta})} \\
&= \sum_{\boldsymbol{A}^*} \frac{\nabla_{\boldsymbol{\Theta}} p(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta})}{p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\Theta})} \\
&= \sum_{\boldsymbol{A}^*} \frac{p(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta})}{p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\Theta})} \frac{\nabla_{\boldsymbol{\Theta}} p(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta})}{p(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta})} \\
&= \sum_{\boldsymbol{A}^*} p(\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta}) \boldsymbol{S}_{comp}(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta}) \\
&= \mathbb{E}_{\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta}} \left[\boldsymbol{S}_{comp}(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta}) \right].
\end{aligned} \tag{B.14}$$

The Fisher Information Matrix (FIM) of the complete data is defined as the covariance of the complete data scores:

$$\begin{aligned}
\mathscr{I}_{comp}(\boldsymbol{\Theta}) &= \mathbb{E}_{\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta}} \left[\boldsymbol{S}_{comp}(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta}) \boldsymbol{S}_{comp}(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta})^\top \right] \\
&= \mathrm{Var}_{\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta}} \left[\boldsymbol{S}_{comp}(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta}) \right].
\end{aligned} \tag{B.15}$$

By the law of total variance, we can decompose the complete data FIM (B.15) as

$$\mathscr{I}_{comp}(\boldsymbol{\Theta}) = \mathbb{E}_{\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}}\left[\mathrm{Var}_{\boldsymbol{A}^*|\boldsymbol{\mathcal{D}},\boldsymbol{\Theta}}\left[\boldsymbol{S}_{comp}(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta})\right]\right] + \mathrm{Var}_{\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}}\left[\mathbb{E}_{\boldsymbol{A}^*|\boldsymbol{\mathcal{D}},\boldsymbol{\Theta}}\left[\boldsymbol{S}_{comp}(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta})\right]\right]. \tag{B.16}$$

From (B.14), we see that the second term in (B.16) is precisely the observed data FIM:

$$\mathscr{I}_{obs}(\boldsymbol{\Theta}) = \mathrm{Var}_{\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}}\left[\boldsymbol{S}_{obs}(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\Theta})\right] = \mathrm{Var}_{\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}}\left[\mathbb{E}_{\boldsymbol{A}^*|\boldsymbol{\mathcal{D}},\boldsymbol{\Theta}}\left[\boldsymbol{S}_{comp}(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta})\right]\right]. \tag{B.17}$$

Moreover, the first term in (B.16) is the expected missing FIM, representing the variance of the complete data scores with respect to the posterior distribution of the missing data $\boldsymbol{A}^*$ given the observed data $\boldsymbol{\mathcal{D}}$ and parameters $\boldsymbol{\Theta}$. This is just the missing data FIM:

$$\mathscr{I}_{miss}(\boldsymbol{\Theta}) = \mathbb{E}_{\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}}\left[\mathrm{Var}_{\boldsymbol{A}^*|\boldsymbol{\mathcal{D}},\boldsymbol{\Theta}}\left[\boldsymbol{S}_{comp}(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta})\right]\right]. \tag{B.18}$$

Combining (B.16), (B.17), and (B.18), we obtain the missing information principle (Louis, 1982):

$$\mathscr{I}_{obs}(\boldsymbol{\Theta}) = \mathscr{I}_{comp}(\boldsymbol{\Theta}) - \mathscr{I}_{miss}(\boldsymbol{\Theta}). \tag{B.19}$$

Assuming there are no additional latent variables in the SCM (Figure 1(a))[1], the complete data likelihood (B.11) composed of several independent components (in terms of the parameters). Thus, the complete data FIM (B.15) is block-diagonal with blocks corresponding to each of the modules in (B.11). For example, under (B.11), we have

$$\mathscr{I}_{comp}(\boldsymbol{\Theta}) = \mathrm{diag}\big(\mathscr{I}_{comp}(\boldsymbol{\eta}), \mathscr{I}_{comp}(\boldsymbol{\beta}_Z), \mathscr{I}_{comp}(\boldsymbol{\gamma}), \mathscr{I}_{comp}(\boldsymbol{\theta})\big)$$

where, for example, $\mathscr{I}_{comp}(\boldsymbol{\eta}) = \mathbb{E}_{\boldsymbol{\mathcal{D}},\boldsymbol{A}^*|\boldsymbol{\eta}}\left[-\nabla^2_{\boldsymbol{\eta}} \log p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{\eta})\right]$. That is, had we observed $\boldsymbol{A}^*$, the parameters of each module would be independent and the complete data FIM block-diagonal. However, the missing data FIM (B.18) is not block-diagonal, as the missing network $\boldsymbol{A}^*$ induces dependencies between the different modules. Therefore, from (B.19) the observed data FIM $\mathscr{I}_{obs}$ is not block-diagonal either. For the observed data FIM to be positive definite and therefore nonsingular, the missing data FIM cross-blocks must not be too large compared to the complete data FIM blocks. Namely, the observed data FIM need to be strictly block-diagonal dominant. This leads to the following sufficient conditions for local identifiability of the parameters $\boldsymbol{\Theta}$ at point $\boldsymbol{\Theta}_0$.

Theorem B.2 (Local Identifiability)**.** *Let $\mathscr{I}_{obs,kk}(\boldsymbol{\Theta}_0) = \mathscr{I}_{comp,kk}(\boldsymbol{\Theta}_0) - \mathscr{I}_{miss,kk}(\boldsymbol{\Theta}_0)$ be the observed data FIM in module k at point $\boldsymbol{\Theta}_0$. The parameters $\boldsymbol{\Theta}$ are locally identifiable at point $\boldsymbol{\Theta}_0$ if $\mathscr{I}_{obs}$ is a strictly block-diagonal dominant matrix. That is, if for every module k,*

$$\lambda_{\min}\left(\mathscr{I}_{comp,kk}(\boldsymbol{\Theta}_0) - \mathscr{I}_{miss,kk}(\boldsymbol{\Theta}_0)\right) > \sum_{j \neq k} \|\mathscr{I}_{miss,jk}(\boldsymbol{\Theta}_0)\|_2, \tag{B.20}$$

where $\lambda_{\min}$ denotes the minimum eigenvalue and $\|\cdot\|_2$ is the spectral norm.

Proof We can show it by contradiction. Assume that (B.20) holds, but $\mathscr{I}_{obs} = \mathscr{I}_{obs}(\boldsymbol{\Theta}_0)$ is singular at $\boldsymbol{\Theta}_0$. Then, there exists a non-zero vector $\mathbf{c}$ such that $\mathscr{I}_{obs}\mathbf{c} = \mathbf{0}$. Partition $\mathbf{c}$ into blocks $\mathbf{c} = (\mathbf{c}_1^\top, \ldots, \mathbf{c}_K^\top)^\top$. Thus, for any block k of $\mathscr{I}_{obs}$, we have

$$\mathscr{I}_{obs,kk}\mathbf{c}_k + \sum_{j \neq k} \mathscr{I}_{obs,jk}\mathbf{c}_j = \mathbf{0}.$$

1. If there are any latent variables, the same approach will work but researchers have to model their process.

By the information decomposition (B.19), we can write it as

$$(\mathscr{I}_{comp,kk} - \mathscr{I}_{miss,kk})\,\mathbf{c}_k + \sum_{j\neq k}(-\mathscr{I}_{miss,jk})\,\mathbf{c}_j = \mathbf{0}.$$

Rearranging terms, we obtain

$$(\mathscr{I}_{comp,kk} - \mathscr{I}_{miss,kk})\,\mathbf{c}_k = \sum_{j\neq k}\mathscr{I}_{miss,jk}\mathbf{c}_j.$$

Taking norms on both sides and using the triangle inequality, we get

$$\|(\mathscr{I}_{comp,kk} - \mathscr{I}_{miss,kk})\,\mathbf{c}_k\|_2 = \left\|\sum_{j\neq k}\mathscr{I}_{miss,jk}\mathbf{c}_j\right\|_2 \leq \sum_{j\neq k}\|\mathscr{I}_{miss,jk}\|_2\|\mathbf{c}_j\|_2$$

As $\mathscr{I}_{comp,kk} - \mathscr{I}_{miss,kk}$ is positive definite and symmetric since condition (B.20) holds, then all its eigenvalues are positive, and we have

$$\|(\mathscr{I}_{comp,kk} - \mathscr{I}_{miss,kk})\,\mathbf{c}_k\|_2 \geq \lambda_{\min}\left(\mathscr{I}_{comp,kk} - \mathscr{I}_{miss,kk}\right)\|\mathbf{c}_k\|_2.$$

Take module k with the largest $\|\mathbf{c}_k\|_2$. Combining the above inequalities, we obtain

$$\lambda_{\min}\left(\mathscr{I}_{comp,kk} - \mathscr{I}_{miss,kk}\right)\|\mathbf{c}_k\|_2 \leq \sum_{j\neq k}\|\mathscr{I}_{miss,jk}\|_2\|\mathbf{c}_j\|_2 \leq \sum_{j\neq k}\|\mathscr{I}_{miss,jk}\|_2\|\mathbf{c}_k\|_2.$$

Dividing both sides by $\|\mathbf{c}_k\|_2 > 0$, we get

$$\lambda_{\min}\left(\mathscr{I}_{comp,kk} - \mathscr{I}_{miss,kk}\right) \leq \sum_{j\neq k}\|\mathscr{I}_{miss,jk}\|_2,$$

which contradicts (B.20). Thus, $\mathscr{I}_{obs}$ is nonsingular, and, by Rothenberg (1971, Theorem 1), the parameters $\boldsymbol{\Theta}$ are locally identifiable. ■

We can give some structure and inuition regarding the (j,k)-th block of $\mathscr{I}_{miss}$. Let $S_k(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta})$ be the score function of module k (e.g., outcome, treatment, proxy, or network model) with respect to its parameters. The partial First-order Taylor expansion for S_k with respect to $\boldsymbol{A}^*$ around the posterior mean $\mathbb{E}[\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta}]$ yields

$$S_k(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta}) \approx S_k(\boldsymbol{\mathcal{D}}, \mathbb{E}[\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta}] \mid \boldsymbol{\Theta}) + \nabla_{\boldsymbol{A}^*} S_k(\boldsymbol{\mathcal{D}}, \mathbb{E}[\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta}] \mid \boldsymbol{\Theta})(\boldsymbol{A}^* - \mathbb{E}[\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta}]),$$

where the gradient is with respect to the entries of $\boldsymbol{A}^*$, and should correspond to the discrete differences since $\boldsymbol{A}^*$ is a binary adjacency matrix. Namely, to changes in one edge at a time. For brevity, write $S_k(\boldsymbol{\mathcal{D}}, \boldsymbol{A}^* \mid \boldsymbol{\Theta}) = S_k(\boldsymbol{A}^*)$. Using the above approximation, we can express the posterior covariance between scores of modules j and k as:

$$\mathrm{Cov}_{\boldsymbol{A}^*|\boldsymbol{\mathcal{D}},\boldsymbol{\Theta}}\left(S_j(\boldsymbol{A}^*), S_k(\boldsymbol{A}^*)\right) \approx \nabla_{\boldsymbol{A}^*} S_j(\mathbb{E}[\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta}])\ \mathrm{Cov}_{\boldsymbol{A}^*|\boldsymbol{\mathcal{D}},\boldsymbol{\Theta}}(\boldsymbol{A}^*)\ \nabla_{\boldsymbol{A}^*} S_k(\mathbb{E}[\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta}])^\top,$$

Thus, the (j,k) cross-block of the missing data FIM (B.18) can be approximated as

$$\mathscr{I}_{miss,jk} \approx \mathbb{E}_{\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}}\left[\left(\nabla_{\boldsymbol{A}^*} S_j\right)\Sigma_{\boldsymbol{A}^*|\boldsymbol{\mathcal{D}},\boldsymbol{\Theta}}\left(\nabla_{\boldsymbol{A}^*} S_k\right)^\top\right], \tag{B.21}$$

where $\Sigma_{\boldsymbol{A}^*|\boldsymbol{\mathcal{D}},\boldsymbol{\Theta}} = \mathrm{Cov}_{\boldsymbol{A}^*|\boldsymbol{\mathcal{D}},\boldsymbol{\Theta}}(\boldsymbol{A}^*)$ is the posterior covariance of the latent network. This expression shows that the missing data FIM between modules j and k depends on how sensitively their scores respond to changes in $\boldsymbol{A}^*$ (the gradients), weighted by the uncertainty in $\boldsymbol{A}^*$ given the data (the posterior covariance). Specifically, $\nabla_{\boldsymbol{A}^*} S_k$ captures how much parameter block k (e.g, $\boldsymbol{\eta}$) is sensitive to one edge changes in $\boldsymbol{A}^*$. If two modules have similar sensitivities to $\boldsymbol{A}^*$ for edges with high posterior uncertainty, then their scores will be correlated, leading to larger off-diagonal blocks in $\mathscr{I}_{miss}$. If that is the case for many modules pairs and edges in $\boldsymbol{A}^*$, then the missing data information matrix will have large off-diagonal entries. By (B.17) the observed data FIM is the difference between the block-diagonal complete data FIM and the dense missing data FIM. Therefore, if the off-diagonal blocks of $\mathscr{I}_{miss}$ are large, then the observed data FIM will be close to singular. On the other hand, if the off-diagonal blocks of $\mathscr{I}_{miss}$ are small compared to the diagonal blocks, then the observed data FIM will be diagonally dominant and thus invertible.

For example, if two modules depend on $\boldsymbol{A}^*$ through similar sufficient statistics (e.g., both depend on the degree of nodes), then their score functions will respond similarly to changes in $\boldsymbol{A}^*$, leading to high covariance in their scores and larger off-diagonal blocks in $\mathscr{I}_{miss}$. Conversely, if modules depend on $\boldsymbol{A}^*$ through distinct sufficient statistics, e.g., if the proxy networks module $p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{\gamma})$ depend on individual edges while the module of outcomes $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{\eta})$ depend on aggregated properties (e.g., number of treated neighbors), then their score functions will respond differently to single edge changes in $\boldsymbol{A}^*$, leading to lower covariance in their scores and smaller off-diagonal blocks in $\mathscr{I}_{miss}$. This analysis suggests that identification improves when the modules depend on $\boldsymbol{A}^*$ through distinct structural features, or when the posterior uncertainty in $\boldsymbol{A}^*$ is low.

Appendix C. Sampling From The Posterior

C.1 Non-Causal Proxies

Under the scenario described in Figures 1(c)-(d) of non-causal proxy networks $\mathcal{A}$, we assume there exist latent variables $\boldsymbol{V}$ such that $\mathcal{A} \leftarrow \boldsymbol{V} \rightarrow \boldsymbol{A}^*$ but without a direct path between the proxies $\mathcal{A}$ and the latent network $\boldsymbol{A}^*$.

As in the main text, in the case of latent-based treatment assignment (Figure 1(c)), the treatment assignment model is $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{A}^*, \boldsymbol{\beta})$. In that case, the joint posterior distribution (5) should also include the unobserved $\boldsymbol{V}$ and can be written as

$$\begin{aligned} p(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\beta_V}, \boldsymbol{V}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}) \propto\ & p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta}) p(\boldsymbol{\eta}) \\ & \times p(\boldsymbol{Z} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\beta}) p(\boldsymbol{\beta}) \\ & \times p(\mathcal{A} \mid \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma}) \\ & \times p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ & \times p(\boldsymbol{V} \mid \boldsymbol{\beta_V}) p(\boldsymbol{\beta_V}), \end{aligned} \tag{C.1}$$

where $p(\boldsymbol{V} \mid \boldsymbol{\beta_V})$ is an assumed model for the latent variables (which could often be assumed to be multivariate normal) parameterized by $\boldsymbol{\beta_V}$ with prior $p(\boldsymbol{\beta_V})$. We obtain that the posterior of $\boldsymbol{A}^*$ given the data and the other unknowns (the analog of (7) that also includes

$\boldsymbol{V}$) is

$$\begin{aligned} p(\boldsymbol{A}^* \mid \cdot) &\equiv p(\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{V}) \\ &\propto p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta}) \\ &\times p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\theta}) \\ &\times p(\boldsymbol{Z} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\beta}), \end{aligned} \tag{C.2}$$

which simplifies the Locally Informed Proposals updates as it does not include the proxy network model $p(\mathcal{A} \mid \cdot)$. Thus, the proposed updates for $\boldsymbol{A}^*$ will depend on the proxies $\mathcal{A}$ only through their information about the latent variable $\boldsymbol{V}$ and the joint structure of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ as will be captured in the posterior sampling through the conditional posterior sampling of $\boldsymbol{V}$ (via the continuous kernel in the Block Gibbs algorithm).

For the scenario of non-causal proxies with proxy-based treatment assignment (Figure 1(d)), the posterior is similar to (C.1) but without the treatment assignment model terms $p(\boldsymbol{Z} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\beta})p(\boldsymbol{\beta})$. That is, in this scenario the treatment model is $p(\boldsymbol{Z} \mid \mathcal{A}, \boldsymbol{X}, \boldsymbol{\beta})$ and does not provide information on $\boldsymbol{A}^*$. Therefore, we can marginalize out $\boldsymbol{\beta}$ and $\boldsymbol{Z}$ from (C.1) (similarly to (6)) to obtain the posterior distribution

$$\begin{aligned} p(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{V}}, \boldsymbol{V}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}) \propto\ & p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta}) p(\boldsymbol{\eta}) \\ & \times p(\mathcal{A} \mid \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma}) \\ & \times p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ & \times p(\boldsymbol{V} \mid \boldsymbol{\beta}_{\boldsymbol{V}}) p(\boldsymbol{\beta}_{\boldsymbol{V}}). \end{aligned} \tag{C.3}$$

Therefore, in this case, the posterior of $\boldsymbol{A}^*$ given the data and the other unknowns is

$$p(\boldsymbol{A}^* \mid \cdot) \propto p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta}) \times p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\theta}), \tag{C.4}$$

which is similar to (C.2) but without the treatment model term. As described below, for estimating the causal effects, we marginalize out all other unknowns except $\boldsymbol{A}^*$ and $\boldsymbol{\eta}$ for the posteriors.

C.2 Extensions

We provide details on how to adapt the posterior in two cases: when researchers want to augment the outcome model with propensity score and when there are latent network-outcome confounders $\boldsymbol{U}$.

- Under latent-based treatment assignment, the treatment model is $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{A}^*, \boldsymbol{\beta})$. As mentioned above, this model informs the outcome model only through its effect on $\boldsymbol{A}^*$. Also, under proxy-based treatment assignment, the treatment model $p(\boldsymbol{Z} \mid \mathcal{A}, \boldsymbol{X}, \boldsymbol{\beta})$ does not provide information on $\boldsymbol{A}^*$, and therefore does not affect the outcome model. In both cases, researchers might want to augment the outcome model with propensity scores $p(\boldsymbol{Z} \mid \cdot)$. However, misspecification of the treatment model can adversely affect the estimation of causal effects (Zigler et al., 2013). To mitigate this issue, it is often suggested to remove 'feedback' between the two models to reduce sensitivity to model misspecification (Zigler et al., 2013). That can be achieved by replacing the continuous updates in the Block Gibbs algorithm with samples from a cut-posterior (Bayarri et al.,

2009; Jacob et al., 2017) of the continuous parameters that removed the feedback term between the propensity scores and the outcome model. Such augmentation will not directly affect $\boldsymbol{A}^*$ updates.

- Assume there is unmeasured network-outcome confounding represented by latent variable $\boldsymbol{U}$, and accounting for it is required for the estimation of, for example, the impact of interventions on the interference network structure. In this case, researchers specify a model $p(\boldsymbol{U} \mid \boldsymbol{\beta}_U)$ for parameters $\boldsymbol{\beta}_U$ with prior $p(\boldsymbol{\beta}_U)$ and assume prior independence as before. The network and outcome models now include $\boldsymbol{U}$: $p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{\theta})$ and $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{\eta})$. As $\boldsymbol{U}$ will typically be modeled as a continuous latent variable (e.g., multivariate normal), it can be included with the rest of the continuous latent variables in the Block Gibbs algorithm. Namely, in $\boldsymbol{A}^*$ updates with LIP will include the current state of $\boldsymbol{U}$ as a fixed variable, and in the continuous updates, $\boldsymbol{U}$ will be updated together with the other continuous latent variables and parameters. A similar approach of network-outcome confounding adjustment using latent variables has been proposed by Um et al. (2024), who assumed the interference network is fully observed.

C.3 Block Gibbs Initialization

As explained in the main text, the proposed Block Gibbs algorithm can be initialized by estimators of latent variables obtained from sampling from the 'cut' posterior (Bayarri et al., 2009; Jacob et al., 2017). This can be followed by a further refinement of $\boldsymbol{A}^*$ using several LIP steps, which is found to greatly improve mixing and convergence of the Block Gibbs sampler in practice. We first present the sampling procedure and then discuss the choice of initial values for the Block Gibbs algorithm.

Sampling from the cut-posterior is a form of Bayesian modularization. It offers several advantages (Bayarri et al., 2009). Notably, it severely reduces, and often eliminates, contamination between modules due to misspecification in some modules. This approach also allows for separate model checking and selection of the network models (true and observed) from that of the outcome model. Researchers can first perform model diagnostics on the network modules and subsequently repeat the process for the outcome model, effectively separating the diagnostic process for the interference network and the outcome model. Moreover, by breaking the complex full posterior into manageable parts, modularization provides a computationally tractable approach where direct sampling would be challenging, as evident in our case.

We consider here the scenario of causal proxies with proxy-based treatment assignment (Figure 1(b)) that have posterior (6), but the approach can be adapted to other scenarios with the posteriors (5), (C.1), or (C.3). The posterior (6) can be written as

$$p(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}) = p(\boldsymbol{\eta} \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^*) p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}).$$

This implies that the posterior is composed of two modules: network module $p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}})$ and outcome module $p(\boldsymbol{\eta} \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^*)$. The modules are connected and can thus generate

'feedback' loops between one another. To see that, notice that

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}) &= \int_{\boldsymbol{\eta}} p(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}) d\boldsymbol{\eta} \\ &\propto p(\boldsymbol{\theta}) p(\boldsymbol{\gamma}) p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma}) p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta}) p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}) \\ &\equiv p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \mathcal{A}, \boldsymbol{X}) p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}) \end{aligned}$$

where $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}) = \int_{\boldsymbol{\eta}} p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta}) p(\boldsymbol{\eta}) d\boldsymbol{\eta}$. Namely, the posterior of $(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^*)$ given the observed data $\boldsymbol{\mathcal{D}}$ is proportional to the posterior given $\boldsymbol{X}$ and $\mathcal{A}$ (the "first module" or "network module") multiplied by $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X})$, the feedback term between the network and outcome modules. Accordingly, the outcome module can be written as

$$\begin{aligned} p(\boldsymbol{\eta} \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^*) &= \frac{p(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}})}{p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}})} \\ &\propto \frac{p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta}) p(\boldsymbol{\eta})}{p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X})} \\ &\propto p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta}) p(\boldsymbol{\eta}), \end{aligned}$$

which can be depicted as $p(\boldsymbol{\eta} \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^*) \equiv p(\boldsymbol{\eta} \mid \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X})$. Viewed as two modules with feedback between them, the posterior can thus be represented as

$$p(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}) \propto \underbrace{p(\boldsymbol{\eta} \mid \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X})}_{\text{Outcome module}} \underbrace{p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \mathcal{A}, \boldsymbol{X})}_{\text{Network module}} \underbrace{p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X})}_{\text{Feedback term}}.$$

The cut-posterior that removes the feedback term from the full posterior is

$$p_{cut}(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}) \propto p(\boldsymbol{\eta} \mid \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}) p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \mathcal{A}, \boldsymbol{X})$$

Sampling from the network module. The network module is proportional to

$$p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \mathcal{A}, \boldsymbol{X}) \propto p(\boldsymbol{\theta}) p(\boldsymbol{\gamma}) p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma}) p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta}),$$

and can also be represented as

$$p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \mathcal{A}, \boldsymbol{X}) = p(\boldsymbol{A}^* \mid \mathcal{A}, \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\gamma}) p(\boldsymbol{\theta}, \boldsymbol{\gamma} \mid \mathcal{A}, \boldsymbol{X}),$$

where

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\gamma} \mid \mathcal{A}, \boldsymbol{X}) &= \sum_{\boldsymbol{A}^*} p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \mathcal{A}, \boldsymbol{X}) \\ &= p(\boldsymbol{\theta}) p(\boldsymbol{\gamma}) \sum_{\boldsymbol{A}^*} p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma}) p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta}). \end{aligned}$$

Thus, sampling from the network module can be performed by first sampling $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ from the marginalized model $p(\boldsymbol{\theta}, \boldsymbol{\gamma} \mid \mathcal{A}, \boldsymbol{X})$, and then sampling networks from $p(\boldsymbol{A}^* \mid \mathcal{A}, \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\gamma})$. For example, if both the network generation model $p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta})$ and the observed network model $p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma})$ are conditionally dyad independent, i.e., can be written as

$$\begin{aligned} p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta}) &= \prod_{i>j} \Pr(A^*_{ij} \mid \boldsymbol{X}, \boldsymbol{\theta}) \\ p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{\gamma}) &= \prod_{i>j} \Pr(A_{ij} \mid A^*_{ij}, \boldsymbol{X}, \boldsymbol{\gamma}), \end{aligned}$$

the sampling from the network module will be simplified, as we now show. Examples for $\Pr(A^*_{ij} \mid \cdot)$ include random graphs, SBM, and latent space models. Examples for $\Pr(A_{ij} \mid \cdot)$ are given in Section 3.2. When $\mathcal{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_B)$, if

$$p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma}) = \prod_b \prod_{i>j} \Pr_b(A_{b,ij} \mid A^*_{ij}, \boldsymbol{X}, \boldsymbol{\gamma}_b),$$

then we proceed the same way. For example, the factorization over b conditional on $\boldsymbol{\gamma}$ is feasible when $\boldsymbol{\gamma}$ has a hierarchical prior. Alternatively, in the case of auto-regressive proxy measurements, we can often write

$$p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma}) = p(\boldsymbol{A}_1 \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma}_1) \prod_{b>1} p(\boldsymbol{A}_b \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_{b-1}, \boldsymbol{\gamma}).$$

For ease of presentation, consider the case of a single proxy network ($B = 1$). We obtain,

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\gamma} \mid \mathcal{A}, \boldsymbol{X}) &\propto p(\boldsymbol{\theta}) p(\boldsymbol{\gamma}) \sum_{\boldsymbol{A}^*} \prod_{i>j} \Pr(A_{ij} \mid A^*_{ij}, \boldsymbol{X}, \boldsymbol{\gamma}) \Pr(A^*_{ij} \mid \boldsymbol{X}, \boldsymbol{\theta}) \\ &= p(\boldsymbol{\theta}) p(\boldsymbol{\gamma}) \prod_{i>j} \sum_{k=0,1} \Pr(A_{ij} \mid A^*_{ij} = k, \boldsymbol{X}, \boldsymbol{\gamma}) \Pr(A^*_{ij} = k \mid \boldsymbol{X}, \boldsymbol{\theta}). \end{aligned} \tag{C.5}$$

Sampling from (C.5) is trivial in most probabilistic programming languages by encoding the posterior with the log-sum-exp trick. We essentially have a mixture of Bernoulli random variables A_{ij}, with mixture probabilities $\Pr(A^*_{ij} = k \mid \boldsymbol{X}, \boldsymbol{\theta})$ and Bernoulli probabilities $\Pr(A_{ij} \mid A^*_{ij} = k, \boldsymbol{X}, \boldsymbol{\gamma})$, both for $k = 0, 1$. In the case of multiple proxies $B > 1$, we will obtain a mixture of categorical random variables (the proxies) with the same mixture probabilities and number of categories equal to 2^B in the case that all proxy networks are unweighted. Note that even if $\mathcal{A}$ or $\boldsymbol{A}^*$ are discrete with more than two categories, we can proceed the same way by modifying the mixture probabilities accordingly. Finally, generating samples of $\boldsymbol{A}^*$ is reduced to sampling edges independently from

$$\begin{aligned} p(\boldsymbol{A}^* \mid \mathcal{A}, \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\gamma}) &= \frac{p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma}) p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta})}{\sum_{\boldsymbol{A}^*} p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\gamma}) p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta})} \\ &= \prod_{i>j} \frac{\Pr(A_{ij} \mid A^*_{ij}, \boldsymbol{X}, \boldsymbol{\gamma}) \Pr(A^*_{ij} \mid \boldsymbol{X}, \boldsymbol{\theta})}{\sum_{k=0,1} \Pr(A_{ij} \mid A^*_{ij} = k, \boldsymbol{X}, \boldsymbol{\gamma}) \Pr(A^*_{ij} = k \mid \boldsymbol{X}, \boldsymbol{\theta})}, \end{aligned} \tag{C.6}$$

which can be performed by simple sampling of $N(N-1)/2$ Bernoulli random variables with probabilities

$$\frac{\Pr(A_{ij} \mid A^*_{ij} = 1, \boldsymbol{X}, \boldsymbol{\gamma}) \Pr(A^*_{ij} = 1 \mid \boldsymbol{X}, \boldsymbol{\theta})}{\sum_{k=0,1} \Pr(A_{ij} \mid A^*_{ij} = k, \boldsymbol{X}, \boldsymbol{\gamma}) \Pr(A^*_{ij} = k \mid \boldsymbol{X}, \boldsymbol{\theta})}$$

In the case of non-causal proxies (C.1), the network module will not include the proxy networks model $p(\mathcal{A} \mid \cdot)$. Therefore, sampling $\boldsymbol{A}^*$ from (C.6) is equivalent to sampling networks from the model $p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\theta})$ given covariates $\boldsymbol{X}$, and current values of the latent variables $\boldsymbol{V}$ and the parameters $\boldsymbol{\theta}$.

Sampling from the outcome module. Given multiple $\boldsymbol{A}^*$ draws from the network module, we can use the modified SCM (2) to sample from the network module. Specifically,

under the modified SCM, outcomes depend on $\boldsymbol{A}^*$ only through summarizing values of ϕ_1, ϕ_2, ϕ_3. Let $\boldsymbol{\phi}_{m,i} = \big(\phi_1(\boldsymbol{Z}_{-i}, \boldsymbol{A}^{*,m}), \phi_2(\boldsymbol{X}_{-i}, \boldsymbol{A}^{*,m}), \phi_{3,i}(\boldsymbol{A}^{*,m})\big)$, be the unit-level summarizing values based on the $m-$th posterior draw from the network module. We average $\boldsymbol{\phi}_{m,i}$ across the posterior draws $\widehat{\boldsymbol{\phi}}_i = M^{-1} \sum_{m=1}^{M} \boldsymbol{\phi}_{m,i}$ and then sample $\boldsymbol{\eta}$ from the outcome module

$$p(\boldsymbol{\eta} \mid \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}) \propto p(\boldsymbol{\eta}) p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta})$$

using the plug-in values $\widehat{\boldsymbol{\phi}} = (\widehat{\boldsymbol{\phi}}_1, \ldots, \widehat{\boldsymbol{\phi}}_N)$ in the outcome model (Bayarri et al., 2009). The outcome model often also depends on $\boldsymbol{A}^*$ through paths that are not possible to summarize with a simple function. In that case, we will plug in the outcome module the sampled $\boldsymbol{A}^{*,m}$ with the largest conditional cut-posterior probability,

$$p(\boldsymbol{A}^* \mid \mathcal{A}, \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \propto p(\mathcal{A} \mid \boldsymbol{A}^*, \boldsymbol{X}, \gamma) p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{\theta})$$

or just $p(\boldsymbol{A}^* \mid \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{\theta})$ in the case of non-causal proxies $\mathcal{A}$.

Choosing initial values for Block Gibbs algorithm. For $\boldsymbol{\theta}, \boldsymbol{\gamma}$ we can either take maximum a-posteriori (MAP) or mean-posterior values as initial values. Given these values, we sample multiple networks $\boldsymbol{A}^*$ using (C.6) and choose the network with the highest conditional cut-posterior log-probability as initial value $\boldsymbol{A}_0^*$. From the outcome module, we took the posterior means of $\boldsymbol{\eta}$ as initial values. As the network module does not use any information from the treatment and outcome models, this initialization strategy for $\boldsymbol{A}^*$ can be improved. We do so by running several LIP updates (Algorithm 1) of $\boldsymbol{A}^*$ given the initial values of $\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\eta}$. We take this refined $\boldsymbol{A}^*$ as the initial value for the Block Gibbs algorithm. The initialization strategy can be summarized as

1. Compute MAP or mean-posterior values of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ from the marginalized network-module in the cut-posterior (C.5).

2. Sample multiple $\boldsymbol{A}^*$ using the values of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ with probabilities (C.6). Save $\boldsymbol{A}_0^*$ as the network with the highest log-probability (C.6). Estimate summary network statistics of the outcome model ϕ_1, ϕ_2, ϕ_3 using a plugin estimator of the mean across the network samples.

3. Sample η from the network model using plug-in $\widehat{\phi}_1, \widehat{\phi}_2, \widehat{\phi}_3$ and $\boldsymbol{A}_0^*$ (if necessary), and take mean posterior (or MAP) values of η.

4. Refine $\boldsymbol{A}_0^*$ by running several LIP updates (Algorithm 1) of $\boldsymbol{A}^*$ given $(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\eta})$ from steps 1 and 3. Take the refined $\boldsymbol{A}^*$ as initial value for the Block Gibbs algorithm.

In the numeric illustrations on fully-synthetic data (Section 6.1), we sample $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ and use their MAP values in step 1 using stochastic variational inference (SVI) with Multivariate Normal variational family and ClippedAdam optimizer with learning rate of 5×10^{-4} which we ran for 2×10^4 steps. In step 3, we estimated the outcome model (given fixed network), with NUTS with 1.2×10^4 posterior samples ($3,000$ samples from four chains). We found that SVI for the marginalized network models was faster with sufficient accuracy, whereas NUTS better explore the posterior space for the outcome model. For the refinement in step 4, we ran $L = 2 \times 10^4$ LIP updates with $K = 5$ edge flip proposal in each step. Start to end, the initialization was very fast to run for all setups considered in this paper.

C.4 Posterior Predictive Distribution of Outcomes

Recall that the observed data are $\boldsymbol{\mathcal{D}} = (\boldsymbol{Y}, \boldsymbol{Z}, \mathcal{A}, \boldsymbol{X})$. We saw in Section 5.3 that we can estimate causal effects using the conditional expectations $\mathbb{E}\big[Y_i \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X}, \boldsymbol{A}_t^*, \boldsymbol{\eta}_t\big]$ where $\boldsymbol{A}_t^*, \boldsymbol{\eta}_t$ are posterior samples of the latent interference network and the parameters of the outcome model, respectively.

Often, we do not have an analytic expression for the conditional expectation $\mathbb{E}\big[Y_i \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X}, \boldsymbol{A}_t^*, \boldsymbol{\eta}_t\big]$. In that case, we can approximate it. The posterior predictive distribution (PPD) of the outcome $\boldsymbol{Y}$ given a new treatment vector $\widetilde{\boldsymbol{Z}}$ and the observed data $\boldsymbol{\mathcal{D}}$ is

$$p(\widetilde{\boldsymbol{Y}} \mid \widetilde{\boldsymbol{Z}}, \boldsymbol{\mathcal{D}}) = \sum_{\boldsymbol{A}^*} \int_{\boldsymbol{\eta}} p(\widetilde{\boldsymbol{Y}} \mid \widetilde{\boldsymbol{Z}}, \boldsymbol{A}^*, \boldsymbol{X}, \boldsymbol{\eta}) p(\boldsymbol{A}^*, \boldsymbol{\eta} \mid \boldsymbol{\mathcal{D}}) d\boldsymbol{\eta},$$

where $p(\boldsymbol{A}^*, \boldsymbol{\eta} \mid \boldsymbol{\mathcal{D}}) = \int_{\boldsymbol{\theta},\boldsymbol{\gamma}} p(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}) d\boldsymbol{\theta} d\boldsymbol{\gamma}$ is the posterior distribution of $\boldsymbol{A}^*$ and $\boldsymbol{\eta}$. Given each posterior sample $\boldsymbol{A}_t^*, \boldsymbol{\eta}_t$ for $t = 1, \ldots, T$ we can draw $S > 0$ outcomes vectors $\boldsymbol{Y}^{(1,t)}, \ldots, \boldsymbol{Y}^{(S,t)}$ with $\boldsymbol{Y}^{(s,t)} = (Y_1^{(s,t)}, \ldots, Y_N^{(s,t)})$ from the outcome model $p(\widetilde{\boldsymbol{Y}} \mid \widetilde{\boldsymbol{Z}}, \boldsymbol{A}_t^*, \boldsymbol{X}, \boldsymbol{\eta}_t)$. These draws can approximate the conditional expectation via

$$\mathbb{E}\big[Y_i \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X}, \boldsymbol{A}_t^*, \boldsymbol{\eta}_t\big] \approx S^{-1} \sum_{s=1}^{S} Y_i^{(s,t)}.$$

Therefore, we can approximate the mean expected outcome of unit i using draws from the PPD

$$\widehat{\mu}_i(\boldsymbol{z} \mid \boldsymbol{\mathcal{D}}) = \mathbb{E}\big[Y_i(\boldsymbol{z}) \mid \boldsymbol{\mathcal{D}}\big] \approx (ST)^{-1} \sum_{t=1}^{T} \sum_{s=1}^{S} Y_i^{(s,t)}.$$

As described in estimation scheme developed in Section 5.3, we can use $\widehat{\mu}_i(\boldsymbol{z} \mid \boldsymbol{\mathcal{D}})$ to estimate the population-level estimands, e.g., $\tau(\boldsymbol{z}, \boldsymbol{z}')$.

In the case of Figure 1(d), where conditioning on $\boldsymbol{A}^*$ opens a back-door path between $\boldsymbol{Z}$ and $\boldsymbol{Y}$, an additional conditioning on the proxies $\mathcal{A}$ is required for identification. As stated in Section 2.3, the back-door adjustment is therefore $\mathbb{E}[Y_i \mid \boldsymbol{Z} = \boldsymbol{z}, \mathcal{A}, \boldsymbol{A}^*, \boldsymbol{X}]$. In that case, the outcome model should be conditioned on $\mathcal{A}$ as well, and the draws from its posterior predictive distribution will approximate the modified conditional expectation.

In the numerical illustrations, we used $S = 1$ and found that for large T (number of posterior samples), this approximation was accurate. That is, for each posterior sample of $\boldsymbol{\eta}$ and $\boldsymbol{A}^*$, we sampled one outcome vector.

C.5 Gradient-Based Approximation of Locally Informed Proposals

Let $\log p(\boldsymbol{A}^* \mid \boldsymbol{\mathcal{D}}, \boldsymbol{\Theta})$ denote the log conditional posterior of the latent network $\boldsymbol{A}^*$ given the data and parameters, as defined in (7). Let $\boldsymbol{A}^{*'} \in H(\boldsymbol{A}^*)$ be a network in the Hamming ball of size one, meaning that $\boldsymbol{A}^{*'}$ differ from $\boldsymbol{A}^*$ in only one coordinate (i.e., one edge flip). Denote the difference in the log-conditional posterior by

$$\Delta(\boldsymbol{A}^{*'}, \boldsymbol{A}^*) = \log p(\boldsymbol{A}^{*'} \mid \cdot) - \log p(\boldsymbol{A}^* \mid \cdot).$$

Assume balancing function $g(a) = \sqrt{a}$ in (8). The *exact* LIP kernel is (9)

$$Q(\boldsymbol{A}^{*'} \mid \boldsymbol{A}^*) \propto \exp\left(\frac{1}{2}\Delta(\boldsymbol{A}^{*'}, \boldsymbol{A}^*)\right),$$

subject to $\boldsymbol{A}^{*'}$ being in the Hamming ball of size one, as mentioned above. We approximate the differences using a first-order Taylor expansion. Let $\tilde{\Delta}(\boldsymbol{A}^{*'}, \boldsymbol{A}^*)$ be the gradient-based approximation of the differences Δ, as defined in (11). Write the *approximate* LIP kernel is

$$Q^{\nabla}(\boldsymbol{A}^{*'} \mid \boldsymbol{A}^*) \propto \exp\left(\frac{1}{2}\tilde{\Delta}(\boldsymbol{A}^{*'}, \boldsymbol{A}^*)\right).$$

We analyze the efficiency loss introduced by using Q^{∇} instead of Q by leveraging theoretical results from Zanella (2020) and Grathwohl et al. (2021). Zanella (2020) established that LIPs with locally balanced functions are asymptotically optimal in terms of Peskun ordering. Specifically, for any function ψ, define the asymptotic variance by

$$\mathrm{Var}_p(\psi, Q) = \lim_{T\to\infty} \frac{1}{T}\mathrm{Var}\left(\sum_{t=1}^{T} \psi(\boldsymbol{A}_t^*)\right),$$

where $\boldsymbol{A}_t^*$ are the samples from proposal Q which is a $p = p(\boldsymbol{A}^* \mid \cdot)$ stationary Markov transition kernel. The smaller $\mathrm{Var}_p(\psi, Q)$, the more efficient the corresponding MCMC algorithm is in estimating the expectation of ψ under p. Second, define the spectral of the Markov transition kernel Q by

$$\mathrm{Gap}(Q) = 1 - \lambda_2,$$

where λ_2 is the second largest eigenvalue of Q, and always satisfies $\mathrm{Gap}(Q) \geq 0$ (Zanella, 2020). The value of $\mathrm{Gap}(Q)$ is closely related to the convergence properties of Q, where values close to 0 indicate slow convergence while large values correspond to fast convergence (Levin and Peres, 2017). Zanella (2020) showed that LIP with locally balanced function minimizes the variance $\mathrm{Var}_p(\psi, Q)$ and maximize the spectral gap $\mathrm{Gap}(Q)$.

Grathwohl et al. (2021, Theorem 1) quantified the performance gap when replacing exact differences with gradient approximations. Assuming the gradient of the log-posterior is L-Lipschitz, they showed

(a) $Var_p(\psi, Q^{\nabla}) \leq \frac{Var_p(\psi,Q)}{c} + \frac{1-c}{c} Var_p(\psi)$

(b) $\mathrm{Gap}(Q^{\nabla}) \geq c\mathrm{Gap}(Q)$

where $c = \exp(-\frac{1}{2}L)$. That is, using the gradient-based approximation is $c \in (0, 1]$ times "more efficient" than the exact differences (Zanella, 2020). Thus, c values closer to 1 imply a minimal efficiency loss due to the approximations. While Grathwohl et al. (2021) derived L values based on the global spectral norm of the Hessian, we refine this bound for our specific setting. Since our proposal is restricted to single-edge flips, the approximation error is governed by the coordinate-wise curvature rather than the global curvature.

Using the Lagrange remainder form of the Taylor expansion, the difference can be written as

$$\Delta(\boldsymbol{A}^{*'}, \boldsymbol{A}^*) = \log p(\boldsymbol{A}^{*'} \mid \cdot) - \log p(\boldsymbol{A}^* \mid \cdot) = \nabla \log p(\boldsymbol{A}^* \mid \cdot)^{\top}\delta + \frac{1}{2}\delta^{\top}\mathcal{H}\delta,$$

where δ is a one-hot vector representing the flipped edge in the upper triangle of $\boldsymbol{A}^*$, and $\mathcal{H}$ is the Hessian matrix at some point between $\boldsymbol{A}^*$ and $\boldsymbol{A}^{*'}$. The approximation error is strictly the quadratic term

$$\text{Error} = \left|\Delta(\boldsymbol{A}^{*'}, \boldsymbol{A}^*) - \nabla \log p(\boldsymbol{A}^* \mid \cdot)^\top \delta\right| = \left|\frac{1}{2}\delta^\top \mathcal{H} \delta\right|. \tag{C.7}$$

Since δ has only one non-zero element, we can bound the error using the diagonal of the Hessian. Let $e = (i, j)$ be the flipped edge, such that $\delta_e = 1$ and the rest are zero. Then $\delta^\top \mathcal{H} \delta = \mathcal{H}_{e,e}$ is the $e, e = ij, ij$ diagonal element of the Hessian, where $\mathcal{H}_{e,e} = \mathcal{H}_{ij,ij} = \frac{\partial^2 \log p(\boldsymbol{A}^* \mid \cdot)}{(\partial A^*_{ij})^2}$. Therefore, we can bound the error in (C.7) by

$$\text{Error} \leq \frac{1}{2} \max_e |\mathcal{H}_{e,e}|.$$

Thus, we can replace the global Lipschitz constant L with the maximum absolute diagonal element of the Hessian

$$c = \exp\left(-\frac{1}{2} \max_e |\mathcal{H}_{e,e}|\right). \tag{C.8}$$

Derivation for the Simplified Model. We verify this constant for the simplified model (4) introduced in Section 4, extended here to include heterogeneous prior network and proxy measurements models:

$$\begin{aligned} A^*_{ij} &\sim \text{Ber}(\theta_{ij}), \\ A_{ij} \mid A^*_{ij} = k &\sim \text{Ber}(\gamma_{k,ij}),\ k = 0, 1, \\ Z_i &\sim \text{Ber}(p_z), \\ Y_i &= \eta_1 Z_i + \eta_2 \sum_{j \neq i} A^*_{ij} Z_j + \varepsilon_i, \end{aligned}$$

where ε_i are iid $N(0, \sigma^2)$. The log-posterior decomposes into the outcome likelihood, proxy likelihood, and network prior

$$\log p(\boldsymbol{A}^* \mid \cdot) = \log p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{\eta}) + \log p(\boldsymbol{A} \mid \boldsymbol{A}^*, \boldsymbol{\gamma}) + \log p(\boldsymbol{A}^* \mid \boldsymbol{\theta}).$$

Let $R_i = Y_i - \eta_1 Z_i$. We have

$$\begin{aligned} \log p(\boldsymbol{A}^* \mid \boldsymbol{\theta}) &= \sum_{i<j} A^*_{ij} \log(\theta_{ij}) + (1 - A^*_{ij}) \log(1 - \theta_{ij}) \\ \log p(\boldsymbol{A} \mid \boldsymbol{A}^*, \boldsymbol{\gamma}) &= \sum_{i<j} A^*_{ij} \left[A_{ij} \log(\gamma_{1,ij}) + (1 - A_{ij}) \log(1 - \gamma_{1,ij})\right] \\ &\quad + \sum_{i<j} (1 - A^*_{ij}) \left[A_{ij} \log(\gamma_{0,ij}) + (1 - A_{ij}) \log(1 - \gamma_{0,ij})\right] \\ \log p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{A}^*, \boldsymbol{\eta}) &= C - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(R_i - \eta_2 \sum_{j \neq i} A^*_{ij} Z_j\right)^2, \end{aligned}$$

where C is some constant. As both $\log p(\boldsymbol{A}^* \mid \boldsymbol{\theta})$ and $\log p(\boldsymbol{A} \mid \boldsymbol{A}^*, \boldsymbol{\gamma})$ are linear in A^*_{ij}, the curvature will be determined only by the outcome model. Recall that $A^*_{ij} = A^*_{ji}$ as the network is assumed to be symmetric. Thus, the gradient with respect to edge $e = ij$ is

$$\nabla_{ij} \log p(\boldsymbol{A}^* \mid \cdot) = \frac{\partial \log p(\boldsymbol{A}^* \mid \cdot)}{\partial A^*_{ij}} = \widetilde{C} + \frac{\eta_2}{\sigma^2} \left[\left(R_i - \eta_2 \sum_{k \neq i} A_{ki} Z_k \right) Z_j + \left(R_j - \eta_2 \sum_{k \neq j} A_{kj} Z_k \right) Z_i \right],$$

for constant $\widetilde{C}$. Therefore, the $e, e = ij, ij$ element on the diagonal of the Hessian is

$$\mathcal{H}_{e,e} = \frac{\partial^2 \log p(\boldsymbol{A}^* \mid \cdot)}{(\partial A^*_{ij})^2} = -\frac{\eta_2^2}{\sigma^2} \left(Z_j^2 + Z_i^2 \right) = -\frac{\eta_2^2}{\sigma^2} \left(Z_j + Z_i \right),$$

as Z_i are binary. Clearly, $|\mathcal{H}_{e,e}|$ is maximized when $Z_i = Z_j = 1$. Substituting this into (C.8) gives the efficiency factor

$$c = \exp\left(-\frac{\eta_2^2}{\sigma^2} \right).$$

This result indicates that the efficiency of the gradient-based approximation depends on the signal-to-noise ratio of the interference effect, but importantly, does not degrade with the network size N. For moderate interference-to-noise ratio with $\eta_2^2 = 0.5\sigma^2$, we have $c \approx 0.6$, suggesting the approximation does not severely degrade the asymptotic efficiency of the Markov chain.

Appendix D. Numerical Illustrations

For the continuous kernels in the Block Gibbs algorithm we used NumPyro (Phan et al., 2019). Code is available at `https://github.com/barwein/BayesProxyNets`.

The prior specification is as follows in all models and setups. $\theta \sim N(0, 3^2)$, $\gamma \sim N(0, 3^2), \eta \sim N(0, 3^2)$, $\sigma \sim \text{HalfNormal}(0, 3^2)$. For the continuous relaxation using the CONCRETE distribution (Maddison et al., 2017), we assumed a temperature equal to 0.5, implying a continuous density in $[0, 1]$ with a U-shape (see Figure 3(b) in Maddison et al., 2017).

Figure D.1 shows how the first-order Taylor approximations using gradients correctly approximate, up to a scale, the manual differences for the Locally Informed Proposals for $\boldsymbol{A}^*$ updates in the fully-synthetic data setting.

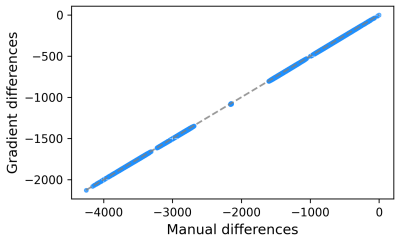


Figure D.1: Gradient approximations $\widetilde{\Delta}\left(ij, t \mid \cdot\right)$ versus the differences $\Delta\left(ij, t \mid \cdot\right)$. Values are on the log-softmax scale. The approximation is accurate up to a scale.

D.1 Fully-Synthetic Data

The full specification of the data generating process for $N = 500$ units is

$$
\begin{aligned}
X_{1,i} &\sim N(0,1) \\
X_{2,i} &\sim Ber(0.1) \\
\Pr(A^*_{ij} = 1 \mid \boldsymbol{X}, \boldsymbol{\theta}) &= s\big(-2 + \widetilde{X}_{2,ij}\big) \\
\Pr(A_{ij} = 1 \mid A^*_{ij}, \boldsymbol{X}, \boldsymbol{\gamma}) &= s\big(A^*_{ij}\gamma_0 + (1 - A^*_{ij})(\gamma_1 + \gamma_2\widetilde{X}_{1,ij} + \gamma_3\widetilde{X}_{2,ij})\big) \\
\Pr(Z_i = 1) &= p_z \frac{d^*_i}{N-1} \\
\varepsilon_i &\sim N(0,1) \\
\mu_i &= -1 + 3Z_i + 2\phi_1 - 0.5X_{1,i} \\
Y_i &= \mu_i + \varepsilon_i,
\end{aligned}
$$

where we set $\gamma_0 = \text{logit}(0.95) - \gamma_2/2$, $\gamma_1 = \text{logit}(0.05) + \gamma_2/2$, and took $\gamma_2 \in \{2, 2.5, 3, 3.5, 4\}$. In each iteration, we resampled all the data. Given each $\boldsymbol{A}^*$ sample, a proxy network is generated for each γ_2 value.

In the BG sampler with one proxy network and HMC for the continuous kernel ("BG (HMC)"), we have used the default settings of HMC, as implemented in NumPyro. That is, a step size of one, a target accept probability of 0.8, and a trajectory length of 2π.

Figure D.2 shows the degree distribution of $\boldsymbol{A}^*$ in one iteration. The right heavy-tails of degrees are units with an extremely high number of connections in the network, representing highly connected units.

Figure D.3 shows the mean ($\pm$ SD) of the Relative Error $\left|\frac{\hat{\tau}-\tau}{\tau}\right|$ of the population-level estimand τ across 300 iterations. The figure is similar to the results of the unit-level MAPE values presented in Figure 3, showing that all BG methods had minimal error, while using a misspecified model ("BG (misspec)") slightly increased the relative error. The continuous relaxation method had a small mean relative error, although with a very high standard deviation across the simulation iterations, highlighting its instability.

Figure D.4 shows the mean ($\pm$ SD) of Relative RMSE values across 300 iterations. Relative RMSE over T posterior samples is defined as $\text{rRMSE} = \sqrt{T^{-1}\sum_{t=1}^{T}\left(\frac{\hat{\tau}_t-\tau}{\tau}\right)^2}$,

where $\hat{\tau}_t$ is point estimate of the estimand at posterior sample t, and τ is the true estimand. Lower values of rRMSE are better.

Figure D.5 presents the frequentist coverage rate in 300 iterations. Coverage is the proportion of iterations where the 95% credible interval included the true estimand τ. Figure D.5 show that the BG algorithm (with one or two proxies) achieves a nominal coverage rate.

Figure D.6 shows traceplots of 1.2×10^4 MCMC samples of the BG sampler in one iteration. The traceplots corresponds to γ_1, η (multiplying ϕ_1), and the dynamic and static estimands. The chains appear to mix well, with no apparent trends or drifts.

Figure D.7 shows the scaling and mixing of the BG across 10 iterations in one setup ($\gamma_2 = 3$ and one proxy network). The figure shows, for varying sample sizes N, the average ($\pm$ SD) of wall-clock time (in seconds) to obtain 1.2×10^4 MCMC samples, and the relative minimal effective sample size (ESS) per second compared to $N = 100$. Even though the running time increases with N, for $N \geq 500$ the relative min ESS/sec stabilizes, indicating that increasing the sample size increases both computation time and sampling efficiency simultaneously.

To evaluate the robustness of the BG sampler when the proxy is, in fact, an accurate measurement of the latent network (i.e., $\boldsymbol{A} \approx \boldsymbol{A}^*$), we performed an additional simulation study in the single-proxy regime ($B = 1$). Using the same DGP described above, we set $\gamma_0 = 15$, $\gamma_1 = -15$, $\gamma_2 = \gamma_3 = 0$, so that $\Pr(A_{ij} = 1 \mid A^*_{ij} = 1) \approx 1 - 3 \times 10^{-7}$ and $\Pr(A_{ij} = 1 \mid A^*_{ij} = 0) \approx 3 \times 10^{-7}$. Consequently, $\boldsymbol{A} = \boldsymbol{A}^*$ with high probability. Critically, the BG sampler uses the same priors on $\boldsymbol{\gamma}$ as in the noisy-proxy settings and is not informed that the proxy is noiseless. Table D.1 reports the mean and standard deviation of the relative error, MAPE, relative RMSE, and posterior standard deviation for the Dynamic and TTE estimands over 100 iterations, comparing BG to the oracle that treats $\boldsymbol{A}^*$ as fixed ("True") and therefore only estimates the outcome-model parameters. The average relative error of BG remains close to zero, indicating that no systematic bias is introduced by modeling $\boldsymbol{A}^*$ as latent. The cost relative to the oracle appears almost entirely in the variance: the posterior standard deviation of the estimands inflates by a factor of roughly 2.55–2.68, and the across-iteration variability of MAPE and relative RMSE grows by a similar order of magnitude. This is the standard efficiency–robustness trade-off of latent-variable models: because BG does not condition on $\boldsymbol{A}^*$, it must estimate $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$, and the network itself, thus propagating the posterior uncertainty about $\boldsymbol{A}^*$ into the causal estimates.

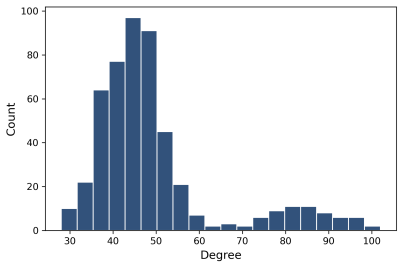


Figure D.2: Degree distribution of $\boldsymbol{A}^*$ in one iteration.

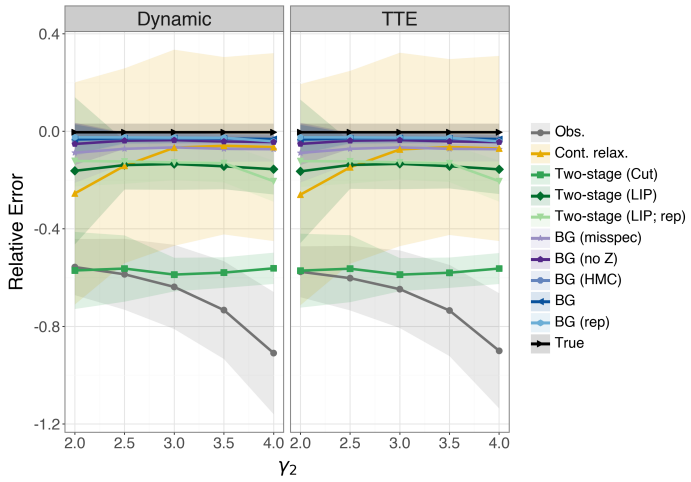


Figure D.3: Mean ($\pm$ SD) of Relative Error $\left|\frac{\hat{\tau}-\tau}{\tau}\right|$ values across 300 iterations.

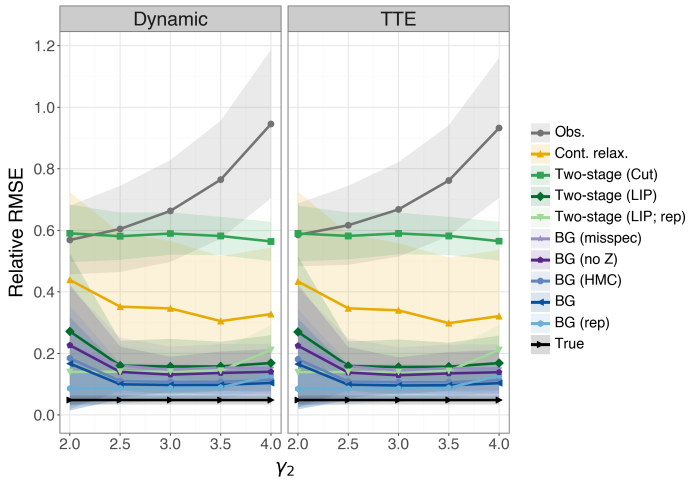


Figure D.4: Mean ($\pm$ SD) of Relative RMSE values across 300 iterations.

Estimand	Method	Relative Error	MAPE	Relative RMSE	Posterior SD
Dynamic	BG	-0.0226 (0.1408)	0.1304 (0.2556)	0.1074 (0.1884)	0.4122 (0.2577)
Dynamic	True	0.0032 (0.0405)	0.0334 (0.0296)	0.0533 (0.0203)	0.1619 (0.0138)
TTE	BG	-0.0392 (0.1705)	0.1077 (0.2125)	0.1085 (0.1891)	1.2875 (0.7907)
TTE	True	0.0032 (0.0407)	0.0289 (0.0283)	0.0510 (0.0209)	0.4798 (0.0313)

Table D.1: Simulation summary of BG sampler using a single proxy network that is equal with high probability to the true network compared to the oracle that takes the true network as a fixed variable. Values are reported as mean (SD) over 100 iterations. "Posterior SD" refers to the posterior standard deviation of the estimand in each iteration.

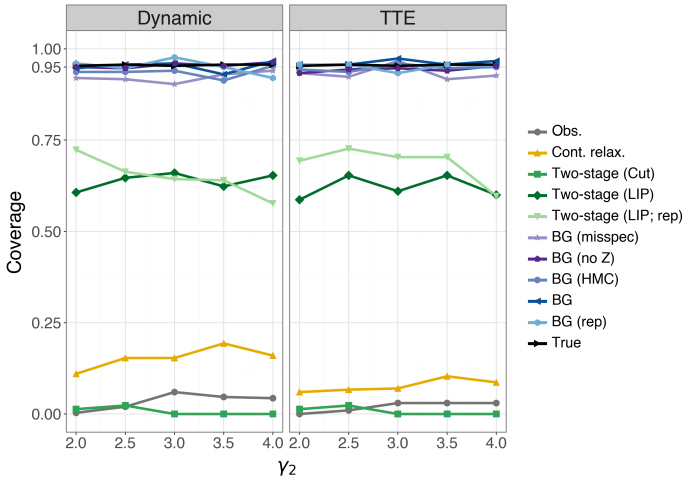


Figure D.5: Coverage rate over 300 iterations.

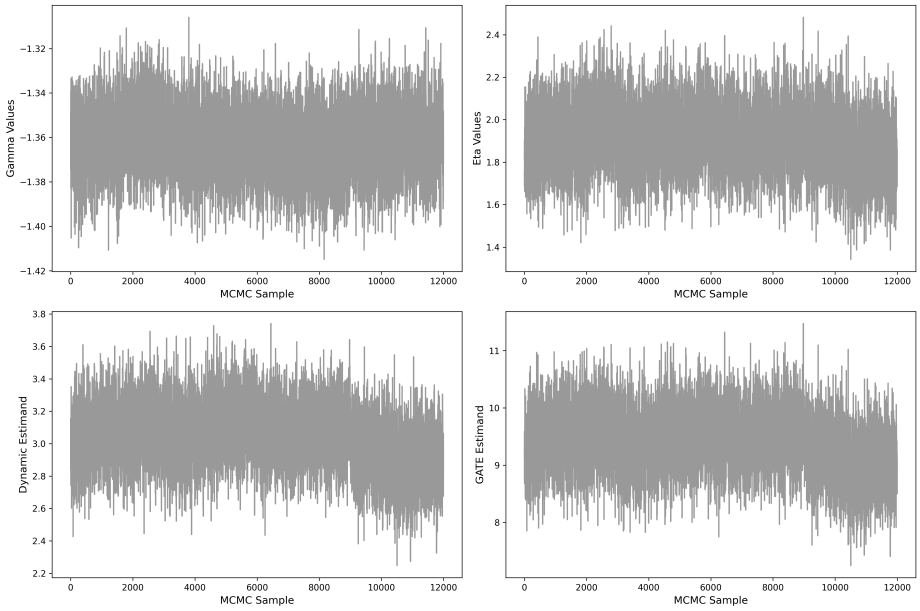


Figure D.6: Traceplots of 1.2×10^4 MCMC samples of the BG sampler in one iteration. The traceplots corresponds to γ_1, η (multiplying ϕ_1), and the dynamic and static estimands.

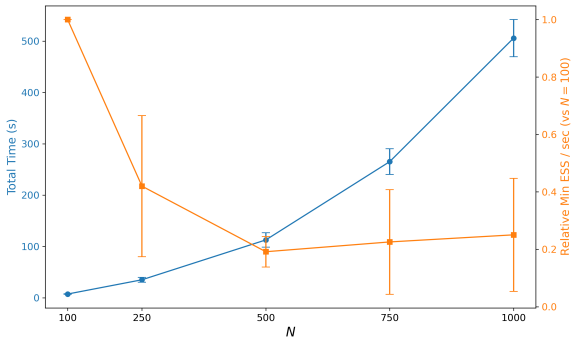


Figure D.7: Scaling and mixing of the BG across 10 iterations in one setup. For varying sample sizes N, the figure shows the average ($\pm$ SD) of wall-clock time (in seconds) to obtain 1.2×10^4 MCMC samples (left y-axis, blue), and the relative minimal effective sample size (ESS) per second compared to $N = 100$ (right y-axis, orange). Even though the running time increase with N, for $N \geq 500$ the relative min ESS/sec stabilizes, indicating that increasing the sample size increase both computation time and sampling efficiency simultaneously. Clock-wall times were obtained on a PC equipped with a 12-core Apple Silicon (ARM64) processor and 48 GB of RAM, running macOS (Darwin Kernel 24.6.0).

D.2 Semi-Synthetic Data

Figure D.8 summarizes the mean (95% posterior interval) of the error $\hat{\tau} - \tau$ in estimating the TTE estimand in each of the setups. The results show that the Block Gibbs sampler had errors close to zero across most iterations and network layers. The Two-stage method had smaller errors than all aggregated networks, other than in the Work layer where the "OR" aggregation had slightly smaller errors. Table D.2 show the empirical coverage of the TTE estimand across 300 iterations by layer and estimation method. The results shows that the BG method achieved nominal coverage across all layers. However, both the Two-stage and the aggregated networks had poor coverage in most layers, with "AND" network being the worst across all layers other than Facebook.

Figure D.9 displays the four networks used in the analysis. Network data is available at `https://manliodedomenico.com/data.php`.

Method	*Facebook*	*Leisure*	*Lunch*	*Work*
True	0.953	0.933	0.957	0.960
BG	0.923	0.933	0.947	0.917
Two-stage	0.577	0.913	0.853	0.854
OR	0.450	0.900	0.607	0.964
AND	0.717	0.545	0.000	0.003

Table D.2: Empricial coverage across 300 iterations of the TTE estimand by layer and estimation method.

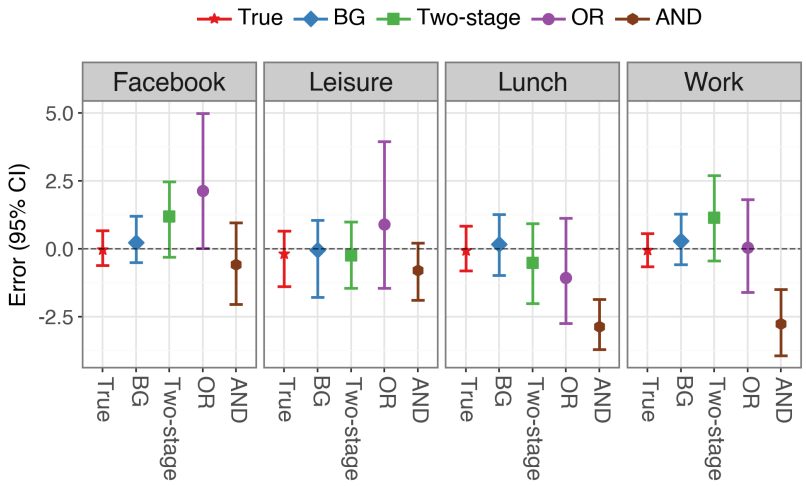


Figure D.8: Mean (95% posterior interval) of error $\hat{\tau} - \tau$ values across 300 iterations in each setup.

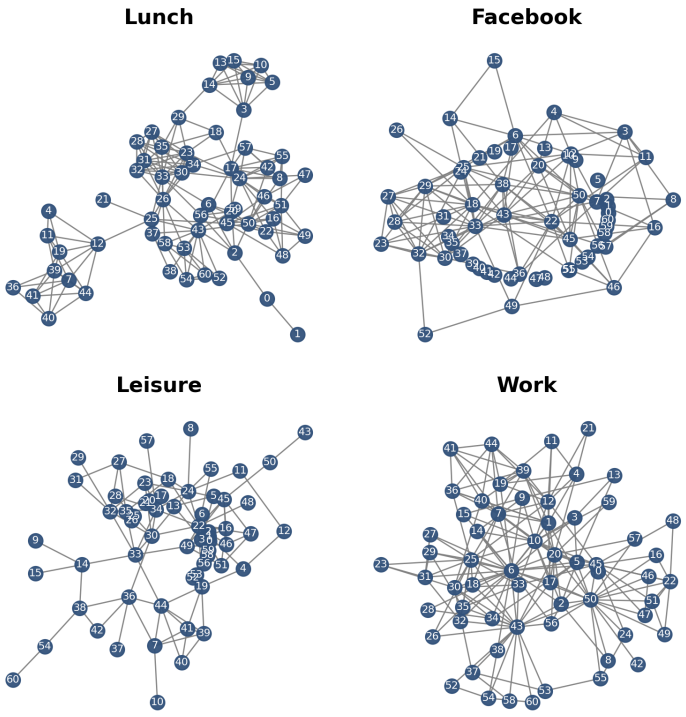


Figure D.9: Multilayer networks.

References

Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37 (6A):3099 – 3132, 2009. doi: 10.1214/09-AOS689.

Peter M. Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4), December 2017. ISSN 1932-6157. doi: 10.1214/16-AOAS1005.

Susan Athey, Dean Eckles, and Guido W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, nov 2018. doi: 10.1080/01621459.2016.1241178.

M. J. Bayarri, J. O. Berger, and F. Liu. Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150, March 2009. ISSN 1936-0975, 1931-6690. doi: 10.1214/09-BA404. Publisher: International Society for Bayesian Analysis.

Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2018.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 2016.

Vincent Boucher and Elysée Aristide Houndetoungan. Estimating peer effects using partial network data. *Centre de recherche sur les risques les enjeux économiques et les politiques*, 2022.

Carter T. Butts. Network inference, error, and informant (in)accuracy: a Bayesian approach. *Social Networks*, 25(2):103–140, May 2003. ISSN 0378-8733. doi: 10.1016/S0378-8733(02)00038-2.

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

Arun Chandrasekhar and Randall Lewis. Econometrics of sampled networks. *Unpublished manuscript, MIT.[422]*, 2011.

Jinyuan Chang, Eric D Kolaczyk, and Qiwei Yao. Estimation of subgraph densities in noisy networks. *Journal of the American Statistical Association*, 117(537):361–374, 2022.

Duncan A Clark and Mark S Handcock. Causal inference over stochastic networks. *Journal of the Royal Statistical Society Series A: Statistics in Society*, January 2024. ISSN 1467-985X. doi: 10.1093/jrsssa/qnae001.

D. R. Cox. *Planning of experiments*. Wiley, New York, 1958. ISBN 9780471181835.

Caterina De Bacco, Martina Contisciani, Jonathan Cardoso-Silva, Hadiseh Safdari, Gabriela Lima Borges, Diego Baptista, Tracy Sweet, Jean-Gabriel Young, Jeremy Koster, Cody T Ross, et al. Latent network models to account for noisy, multiply reported social network data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(3): 355–375, 2023.

Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1), feb 2017. doi: 10.1515/jci-2015-0021.

Paul Erdős and Alfréd Rényi. On random graphs i. *Publicationes mathematicae*, 6(1): 290–297, 1959.

Stephen E. Fienberg. A Brief History of Statistical Models for Network Analysis and Open Challenges. *Journal of Computational and Graphical Statistics*, 21(4):825–839, October 2012. ISSN 1061-8600, 1537-2715. doi: 10.1080/10618600.2012.738106.

Laura Forastiere, Edoardo M. Airoldi, and Fabrizia Mealli. Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534):901–918, jun 2020. doi: 10.1080/01621459.2020.1768100.

Laura Forastiere, Fabrizia Mealli, Albert Wu, and Edoardo M Airoldi. Estimating causal effects under network interference with bayesian generalized propensity scores. *Journal of Machine Learning Research*, 23(289):1–61, 2022.

Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

Paul Goldsmith-Pinkham and Guido W. Imbens. Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264, 2013. ISSN 07350015.

Ravi Goyal, Nicole Carnegie, Sally Slipher, Philip Turk, Susan J. Little, and Victor De Gruttola. Estimating contact network properties by integrating multiple data sources associated with infectious diseases. *Statistics in Medicine*, 2023. ISSN 1097-0258. doi: 10.1002/sim.9816.

Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, and Chris Maddison. Oops i took a gradient: Scalable sampling for discrete distributions. In *International Conference on Machine Learning*, pages 3831–3841. PMLR, 2021.

Alan Griffith. Name your friends, but only five? the importance of censoring in peer effects estimates using social network data. *Journal of Labor Economics*, nov 2021. doi: 10.1086/717935.

Mark S. Handcock and Krista J. Gile. Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5–25, March 2010.

ISSN 1932-6157, 1941-7330. doi: 10.1214/08-AOAS221. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-4/issue-1/Modeling-social-networks-from-sampled-data/10.1214/08-AOAS221.full. Publisher: Institute of Mathematical Statistics.

Morgan Hardy, Rachel M. Heath, Wesley Lee, and Tyler H. McCormick. Estimating spillovers using imprecisely measured networks, 2019.

Alex Hayes, Mark M Fredrickson, and Keith Levin. Estimating network-mediated causal effects via principal components network regression. *Journal of Machine Learning Research*, 26(13):1–99, 2025.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, dec 2002. doi: 10.1198/016214502388618906.

Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Paul W. Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2287037.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, jun 1983. doi: 10.1016/0378-8733(83)90021-7.

Michael G Hudgens and M. Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, jun 2008. doi: 10.1198/016214508000000292.

Pierre E Jacob, Lawrence M Murray, Chris C Holmes, and Christian P Robert. Better together? statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719*, 2017.

Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.

Eric D. Kolaczyk. *Statistical analysis of network data*. Springer, 2009. ISBN 9780387881454.

Can M Le and Tianxi Li. Linear regression and its inference on noisy network-linked data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1851–1885, 2022.

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

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Fan Li, Peng Ding, and Fabrizia Mealli. Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2247), mar 2023. doi: 10.1098/rsta.2022.0153.

Shuangning Li and Stefan Wager. Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics*, 50(4):2334 – 2358, 2022. doi: 10.1214/22-AOS2191.

Tianxi Li, Elizaveta Levina, and Ji Zhu. Prediction models for network-linked data. *The Annals of Applied Statistics*, 13(1):132 – 164, 2019. doi: 10.1214/18-AOAS1205.

Wenrui Li, Daniel L. Sussman, and Eric D. Kolaczyk. Causal inference under network interference with noise, 2021.

Thomas A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233, 1982.

Yunpu Ma and Volker Tresp. Causal inference under networked interference and intervention policy enhancement. In *International Conference on Artificial Intelligence and Statistics*, pages 3700–3708. PMLR, 2021.

C Maddison, A Mnih, and Y Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of the International Conference on Learning Representations*, 2017.

Matteo Magnani, Barbora Micenkova, and Luca Rossi. Combinatorial analysis of multiple networks. *arXiv preprint arXiv:1303.4986*, 2013.

Charles F. Manski. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531, jul 1993. doi: 10.2307/2298123.

Charles F. Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, feb 2013. doi: 10.1111/j.1368-423x.2012.00368.x.

Peter V Marsden. Network data and measurement. *Annual review of sociology*, 16(1): 435–463, 1990.

Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, and Chris J Oates. Generalized bayesian inference for discrete intractable likelihood. *Journal of the American Statistical Association*, 119(547):2345–2355, 2024.

Tyler H McCormick and Tian Zheng. Latent surface models for networks using aggregated relational data. *Journal of the American Statistical Association*, 110(512):1684–1695, 2015.

Radford M Neal. MCMC using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Akihiko Nishimura, David B Dunson, and Jianfeng Lu. Discontinuous Hamiltonian Monte Carlo for discrete parameters and discontinuous likelihoods. *Biometrika*, 107(2):365–380, June 2020. ISSN 0006-3444. doi: 10.1093/biomet/asz083.

Elizabeth L. Ogburn, Oleg Sofrygin, Iván Díaz, and Mark J. van der Laan. Causal inference for social network data. *Journal of the American Statistical Association*, pages 1–46, 2024. doi: 10.1080/01621459.2022.2131557.

Yuki Ohnishi, Bikram Karmakar, and Arman Sabbaghi. Degree of interference: A general framework for causal inference under interference. *arXiv preprint arXiv:2210.17516*, 2022.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

David Puelz, Guillaume Basse, Avi Feller, and Panos Toulis. A graph-theoretic approach to randomization tests of causal effects under general interference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):174–204, February 2022. ISSN 1369-7412, 1467-9868. doi: 10.1111/rssb.12478.

Steven Wilkins Reeves, Shane Lubold, Arun G Chandrasekhar, and Tyler H McCormick. Model-based inference and experimental design for interference using partial network data. *arXiv preprint arXiv:2406.11940*, 2024.

Thomas J Rothenberg. Identification in parametric models. *Econometrica: Journal of the Econometric Society*, pages 577–591, 1971.

Michael Salter-Townshend and Tyler H McCormick. Latent space models for multiview network data. *The annals of applied statistics*, 11(3):1217, 2017.

Fredrik Sävje. Causal inference with misspecified exposure mappings: separating definitions and assumptions. *Biometrika*, 111(1):1–15, 2024.

Fredrik Sävje, Peter M. Aronow, and Michael G. Hudgens. Average treatment effects in the presence of unknown interference. *The Annals of Statistics*, 49(2), apr 2021. doi: 10.1214/20-aos1973.

Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40 (2):211–239, may 2011. doi: 10.1177/0049124111404820.

Sadegh Shirani and Mohsen Bayati. Causal message-passing for experiments with unknown and general network interference. *Proceedings of the National Academy of Sciences*, 121 (40):e2322232121, 2024.

Oleg Sofrygin and Mark J van der Laan. Semi-parametric estimation and inference for the mean outcome of the single time-point intervention in a causally connected population. *Journal of causal inference*, 5(1):20160003, 2017.

Juan Sosa and Brenda Betancourt. A latent space model for multilayer network data. *Computational Statistics & Data Analysis*, 169:107432, May 2022. ISSN 0167-9473. doi: 10.1016/j.csda.2022.107432.

Stan Development Team. Latent discrete parameters. Stan User's Guide, version 2.35, 2024. URL https://mc-stan.org/docs/stan-users-guide/latent-discrete.html.

Eric J. Tchetgen Tchetgen, Isabel R. Fulcher, and Ilya Shpitser. Auto-g-computation of causal effects on a network. *Journal of the American Statistical Association*, 116(534): 833–844, oct 2020. doi: 10.1080/01621459.2020.1811098.

Panos Toulis and Edward Kao. Estimation of causal peer influence effects. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1489–1497, 17–19 Jun 2013.

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

Seungha Um, Tracy Sweet, and Samrachana Adhikari. A bayesian approach to estimate causal peer influence accounting for latent network homophily. *arXiv preprint arXiv:2405.14789*, 2024.

Bar Weinstein and Daniel Nevo. Causal inference with misspecified network interference structure. *Biometrics*, 82(1):ujag023, 02 2026. ISSN 0006-341X. doi: 10.1093/biomtc/ujag023. URL https://doi.org/10.1093/biomtc/ujag023.

Jean-Gabriel Young, George T. Cantwell, and M. E. J. Newman. Bayesian inference of network structure from unreliable data. *Journal of Complex Networks*, 8(6):cnaa046, March 2021. ISSN 2051-1310, 2051-1329. doi: 10.1093/comnet/cnaa046. arXiv:2008.03334 [physics, stat].

Christina Lee Yu, Edoardo M. Airoldi, Christian Borgs, and Jennifer T. Chayes. Estimating the total treatment effect in randomized experiments with unknown network structure. *Proceedings of the National Academy of Sciences*, 119(44), oct 2022. doi: 10.1073/pnas.2208975119.

Giacomo Zanella. Informed proposals for local mcmc in discrete spaces. *Journal of the American Statistical Association*, 2020.

Yichuan Zhang, Zoubin Ghahramani, Amos J Storkey, and Charles Sutton. Continuous Relaxations for Discrete Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, 2012.

Guangyao Zhou. Mixed Hamiltonian Monte Carlo for Mixed Discrete and Continuous Variables. In *Advances in Neural Information Processing Systems*, 2020.

Xuening Zhu, Rui Pan, Guodong Li, Yuewen Liu, and Hansheng Wang. Network vector autoregression. *The Annals of Statistics*, 45(3):1096 – 1123, 2017. doi: 10.1214/16-AOS1476.

Corwin M Zigler, Krista Watts, Robert W Yeh, Yun Wang, Brent A Coull, and Francesca Dominici. Model feedback in bayesian propensity score estimation. *Biometrics*, 69(1): 263–273, 2013.