Causal Inference Under Interference And Network Uncertainty

Rohit Bhattacharya   **Daniel Malinsky**   **Ilya Shpitser**

Department of Computer Science
Johns Hopkins University
{rbhattacharya@, malinsky@, ilyas@cs.}jhu.edu

Abstract

Classical causal and statistical inference methods typically assume the observed data consists of independent realizations. However, in many applications this assumption is inappropriate due to a network of dependences between units in the data. Methods for estimating causal effects have been developed in the setting where the structure of dependence between units is known exactly [10, 36, 20], but in practice there is often substantial uncertainty about the precise network structure. This is true, for example, in trial data drawn from vulnerable communities where social ties are difficult to query directly. In this paper we combine techniques from the structure learning and interference literatures in causal inference, proposing a general method for estimating causal effects under data dependence when the structure of this dependence is not known a priori. We demonstrate the utility of our method on synthetic datasets which exhibit network dependence.

1   INTRODUCTION

In many scientific and policy settings, research subjects do not exist in isolation but in interacting networks. For instance, data drawn from an online social network will exhibit *homophily* (friends are similar, because they are friends), and *contagion* (friends may causally influence each other) [17, 29, 13, 30]. Similarly, vaccinating some subset of a population may confer immunity to the entire population – a well-documented phenomenon known as *herd immunity* in infectious disease epidemiology. This implies that a treatment given to one unit affects outcomes for another. Finally, resource constraints in allocation problems may also induce data dependence.

In the context of causal inference, methods for dealing with data dependence are developed under the heading of *interference* [32, 9, 27, 10, 36, 29, 20, 35]. Most such work assumes the structure of the dependence (which units depend on which others, and how) is known precisely. For example, [36] assumes units in the data may be organized into equal sized blocks, where units within a block are pairwise dependent and units across blocks are not. Some work makes alternative assumptions, e.g., [35] assumes that blocks are drawn from a known random field.

In many applications, the network inducing dependence between units may not be known exactly. For instance, in vulnerable, stigmatized, or isolated communities (such as groups of drug users, or remote villages), we may have no way of reconstructing the precise social ties between individuals. Some online databases of social media users may be anonymized, with friendship ties deliberately omitted. There has been some work in such settings that involves adapting the data collection method itself in order to discover the underlying networks: e.g., snowball sampling in [5] and [3]. Unfortunately, such study designs are not always possible to arrange in advance, and most data available on networks of interacting units is not collected under such designs.

While there is a rich literature on model selection from observational data in the context of causal inference (e.g., [33, 4, 31, 23]), to our knowledge all previous work has assumed the absence of interference. We explore learning the dependence structure using graphical model selection methods. Techniques for structure learning from probabilistic relational models are also related to this work [19, 16].

The contributions of our paper may be viewed in one of two ways. From the point of view of causal inference under interference, our paper contributes to methods for estimating causal effects when there is substantial uncertainty about network structure. From the point of view

of structure learning, we introduce novel algorithms for model selection when units are dependent due to a network, the structure of which is unknown.

2 MOTIVATING EXAMPLE AND BACKGROUND ASSUMPTIONS

To motivate our work, we discuss an example application. Consider a public health program aimed at lowering the incidence of blood-borne diseases such as HIV in at-risk individuals who are addicted to heroin and share needles when injecting intravenously. An example of such a program is described in [34]. The program creates pop-up clinics around the city where disposable needles are distributed for free to individuals in need, but due to limited resources only a limited number of individuals will actually receive these needles. We would like to know, in this restricted resource setting, if the use of disposable needles spreads amongst the rest of the population. Additionally, we would like to detect the phenomenon of herd immunity, i.e., whether some members of the population being protected due to taking advantage of the clean needles confer this protection to others who do not.

Data on heroin users was collected via such program, with users arranged by neighborhood or municipality. Users in different neighborhoods are assumed independent, but users within the same neighborhood are likely dependent. This setting is known as *partial interference* [10]. For each individual i, data is collected on their use of disposable needles A_i, their subsequent health outcome Y_i (risk of obtaining blood-born disease), along with a vector of pre-treatment covariates $\mathbf{L}_i = (L_{1,i}, ..., L_{p,i})$. We may be interested in quantifying the causal effects of A_i on Y_j, for arbitrary i and j within a neighborhood, or network-averaged versions of such effects [20].

We may assume that background knowledge or study design implies a "known" individual-level causal structure for each i, namely that $A_i \to Y_i$ and $A_i \leftarrow \mathbf{L}_i \to Y_i$, but that we are uncertain about network ties among users. One approach is to assume the least restrictive model, where all users in a neighborhood are arbitrarily dependent. This would correspond to a complete network, where every pair of vertices is directly connected. However, assuming a complete network when the true network is sparse ignores useful structure in the problem and leads to inefficient estimates of target quantities. In addition, complete networks often lead to likelihoods that are intractable to evaluate. An alternative is to a select a sparse network supported by the data. In addition to enabling tractable and statistically efficient inference, such an approach may also rule out the presence of cer-
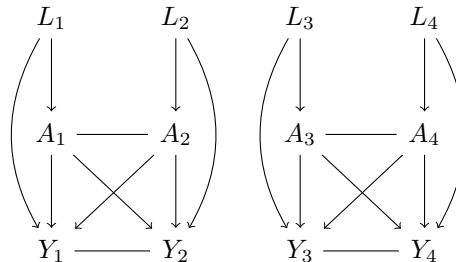

Figure 1: A chain graph over three variables (L, A, and Y) on 4 individuals, representing possible relationships between disposable needle use and risk of blood-borne disease among heroin-users.

tain causal effects without explicitly estimating them, if corresponding pathways are absent in the selected network.

As an example, if neighborhoods have 4 units, we may aim to learn a graphical model such as shown in Figure 1. This model, containing both directed edges (representing direct causal influences) and undirected edges (representing symmetric network ties), is known as a chain graph model [14]. We describe chain graphs in more detail below. This model tells us that we should expect some spread of disposable needle use from one unit to another. However, it also tells us that users in neighborhoods are split into two non-interacting groups: $\{1, 2\}$ and $\{3, 4\}$. This implies the absence of contagion from one group to another. In addition, the conditional independences among units implied by this split suggests that contagion effects within groups may be estimated more efficiently as compared to a statistically saturated model, with a complete network across units.

The algorithms we propose are consistent (in the sense that they asymptotically converge on the true model) under a set of assumptions which we now informally summarize. We assume the true data-generating process corresponds perfectly (satisfying Markov and faithfulness conditions) to some unknown chain graph, with two restrictions: (1) the unit-level graph is known, reflecting the aforementioned causal ordering between pre-treatment covariates, treatment variables, and outcomes; and (2) the graph respects what we later call *tier symmetry*, which restricts connections between variables at the same "tier" in the causal ordering to be symmetric. We assume the data is distributed with some (known) likelihood in the exponential family, as well as some weak statistical regularity conditions. We also present algorithms that make an additional simplifying assumption on the graphical structure – namely that influence between units is the same for all unit pairs – but such an assumption is not strictly necessary for consistency.

We begin by describing some technical preliminaries, including chain graph models, causal inference, and graphical model selection. Then we present algorithms to learn graphical models of the sort shown in Figure 1, before estimating causal effects.

3 PRELIMINARIES

3.1 Graphical Terminology

Chain graphs (CGs) are a class of mixed graphs containing directed ($\rightarrow$) and undirected ($-$) edges, such that it is impossible to create a directed cycle by orienting any combination of the undirected edges [14]. A CG with no undirected edges is a directed acyclic graph (DAG). A CG with no directed edges is an undirected graph (UG) Vertices of a graph are denoted by capital letters (e.g. A), and they correspond to random variables. We use boldface (e.g. $\mathbf{A}$) to denote sets of vertices or sets of random variables. Lowercase letters denote specific values of random variables (e.g. a) or sets of values (e.g. $\mathbf{a}$). We use $\mathbf{V}$ and $\mathcal{E}$ to denote the set of all vertices and edges in a graph $\mathcal{G}$, respectively.

For a subset of vertices $\mathbf{A} \subseteq \mathbf{V}$ we define the induced subgraph $\mathcal{G}_{\mathbf{A}}$ to be the graph with vertices $\mathbf{A}$ and edges of $\mathcal{G}$ that have both endpoints in $\mathbf{A}$. A block $\mathbf{B}$ is defined as a maximal set of vertices such that every vertex pair in $\mathcal{G}_{\mathbf{B}}$ is connected by an undirected path. The set of blocks in a CG $\mathcal{G}$, denoted by $\mathcal{B}(\mathcal{G})$, partitions the vertices in $\mathcal{G}$. A clique $\mathbf{C}$ is defined as a maximal set of vertices that are pairwise connected by undirected edges. A clique in a CG is always a subset of some block $\mathbf{B}$. We denote the set of all cliques in an UG $\mathcal{G}$ by $\mathcal{C}(\mathcal{G})$.

For a graph $\mathcal{G}$ and vertex $V \in \mathbf{V}$ we define some standard vertex sets as follows: the set of parents $\mathrm{pa}_{\mathcal{G}}(V) \equiv \{V' : V' \rightarrow V \text{ in } \mathcal{G}\}$; the set of neighbors $\mathrm{nb}_{\mathcal{G}}(V) \equiv \{V' : V' - V \text{ in } \mathcal{G}\}$; the boundary $\mathrm{bd}_{\mathcal{G}}(V) \equiv \mathrm{pa}_{\mathcal{G}}(V) \cup \mathrm{nb}_{\mathcal{G}}(V)$; and the closure $\mathrm{cl}_{\mathcal{G}}(V) \equiv \mathrm{bd}_{\mathcal{G}}(V) \cup V$. These definitions generalize disjunctively to sets, e.g. $\mathrm{pa}_{\mathcal{G}}(\mathbf{A}) \equiv \bigcup_{A \in \mathbf{A}} \mathrm{pa}_{\mathcal{G}}(A)$. Note that for a block $\mathbf{B}$, $\mathrm{bd}_{\mathcal{G}}(\mathbf{B}) = \mathbf{B} \cup \mathrm{pa}_{\mathcal{G}}(\mathbf{B})$. Given a CG $\mathcal{G}$, define the augmented graph $\mathcal{G}^a$ to be an UG constructed from $\mathcal{G}$ by replacing all directed edges with undirected edges and connecting all vertices in $\mathrm{pa}_{\mathcal{G}}(\mathbf{B})$ for every block $\mathbf{B}$ in $\mathcal{G}$ by undirected edges.

We will utilize chain graphs to represent both causal relationships and network dependence among units that form a ("social") network $\mathcal{N}$. The undirected network $\mathcal{N}$ is a graph (distinct from our CG of interest) where the vertices correspond to units (e.g. individuals $i, j, ...$), not random variables. Units may be adjacent or non-adjacent in $\mathcal{N}$ based on whether they are "friends" or otherwise

directly dependent.

For each unit i, we denote the unit-level variables for i in the CG $\mathcal{G}$ (e.g., L_i, A_i, and Y_i in Figure 1) by $\mathbf{V}_i$, and edges among those variables by $\mathcal{E}_i$. Similarly, for a pair of units i, j which are adjacent in $\mathcal{N}$, we represent the set of edges from $\mathbf{V}_i$ to $\mathbf{V}_j$ (and vice versa) by $\mathcal{E}_{ij}$. It is the presence of these edges that induces data dependence between i and j in our analysis. The set of $\mathcal{E}_{ij}$ for all pairs i, j adjacent in $\mathcal{N}$ (i.e., the set of all cross-unit edges) will be denoted by $\mathcal{E}_{\mathcal{N}}$.

3.2 Chain Graph Models

A statistical chain graph model associated with a LWF (Lauritzen-Wermuth-Frydenberg) chain graph $\mathcal{G}$ is a set of distributions that factorize as:

$$p(\mathbf{V}) = \prod_{\mathbf{B} \in \mathcal{B}(\mathcal{G})} p(\mathbf{B} \mid \mathrm{pa}_{\mathcal{G}}(\mathbf{B}))$$

and $\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ (1)

$$p(\mathbf{B} \mid \mathrm{pa}_{\mathcal{G}}(\mathbf{B})) = \frac{\prod_{\{\mathbf{C} \in \mathcal{C}((\mathcal{G}_{\mathrm{bd}_{\mathcal{G}}(\mathbf{B})})^a) : \mathbf{C} \not\subseteq \mathrm{pa}_{\mathcal{G}}(\mathbf{B})\}} \phi_{\mathbf{C}}(\mathbf{C})}{Z(\mathrm{pa}_{\mathcal{G}}(\mathbf{B}))}$$

for each block $\mathbf{B}$ in $\mathcal{G}$, where $\phi_{\mathbf{C}}(\mathbf{C})$ is a clique potential function for a clique $\mathbf{C}$ in the UG $(\mathcal{G}_{\mathrm{bd}_{\mathcal{G}}(\mathbf{B})})^a$ defined as above and $Z(\mathrm{pa}_{\mathcal{G}}(\mathbf{B}))$ is a normalizing function [14].

A CG without undirected edges is a DAG, which has a simpler factorization: $p(\mathbf{V}) = \prod_{V \in \mathbf{V}} p(V \mid \mathrm{pa}_{\mathcal{G}}(V))$. If it is the case that for every block $\mathbf{B}$ in CG $\mathcal{G}$, $\mathcal{G}_{\mathrm{bd}_{\mathcal{G}}(\mathbf{B})}$ has missing edges only among elements of $\mathrm{pa}_{\mathcal{G}}(\mathbf{B})$, then $(\mathcal{G}_{\mathrm{bd}_{\mathcal{G}}(\mathbf{B})})^a$ has a single clique containing all elements in $\mathrm{bd}_{\mathcal{G}}(\mathbf{B})$. In other words, the model corresponding to such a CG may be viewed as a DAG model with entire blocks $\mathbf{B}$ acting as vertices in a DAG.

3.3 Causal Models

A causal model is a set of distributions over counterfactual random variables, a.k.a. potential outcomes. For $Y \in \mathbf{V}$ and $\mathbf{A} \subseteq \mathbf{V} \setminus Y$, the counterfactual $Y(\mathbf{a})$ denotes the value of Y when the "treatment" variables $\mathbf{A}$ are fixed to values $\mathbf{a}$ by an intervention. Sometimes interventions are formalized by the 'do'-operator: $\mathrm{do}(\mathbf{a})$ denotes the assigment $\mathbf{a}$ to $\mathbf{A}$ [21]. The counterfactual distribution corresponding to the intervention where $\mathbf{A}$ is set to $\mathbf{a}$ is written $p(Y(\mathbf{a}))$ or $p(Y \mid \mathrm{do}(\mathbf{a}))$.

A causal model of a DAG $\mathcal{G}$ is a set of distributions defined on counterfactual random variables $V(\mathbf{a})$ for each $V \in \mathbf{V}$ and where $\mathbf{a}$ is a set of values for $\mathrm{pa}_{\mathcal{G}}(V)$. Equivalently, a causal model can be understood as the set of distributions induced by a system of structural equations (one equation for each vertex) equipped with the $\mathrm{do}(\cdot)$

operator [21, 25]. In a causal model of a DAG $\mathcal{G}$, all counterfactual distributions are *identified*, i.e., they can be expressed as functions of the observed data, by the g-formula [26]:

$$p(\mathbf{V} \setminus \mathbf{A} \mid \mathrm{do}(\mathbf{a})) = \prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V \mid \mathrm{pa}_{\mathcal{G}}(V)) \Bigg|_{\mathbf{A} = \mathbf{a}}.$$

Counterfactual responses to interventions are often contrasted on a mean difference scale under two possible interventions $\mathbf{a}$ and $\mathbf{a}'$, representing cases and controls. For example, the average causal effect (ACE) is given by $\mathbb{E}[Y(\mathbf{a})] - \mathbb{E}[Y(\mathbf{a}')]$.

Causal models have been generalized from DAGs to CGs (details in the Supplement) and yield the following generalization of the g-formula [15]:

$$p(\mathbf{V} \setminus \mathbf{A} \mid \mathrm{do}(\mathbf{a})) = \prod_{\mathbf{B} \in \mathcal{B}(\mathcal{G})} p(\mathbf{B} \setminus \mathbf{A} \mid \mathrm{pa}(\mathbf{B}), \mathbf{B} \cap \mathbf{A}) \Bigg|_{\mathbf{A} = \mathbf{a}}. \tag{2}$$

3.4 The Conditionally Ignorable Network Model and Network Causal Effects

For the purposes of this paper, we consider CGs decomposed into three disjoint sets of variables: $\mathbf{L}$, representing vectors of baseline (pre-treatment) factors; $\mathbf{A}$, representing treatments; and $\mathbf{Y}$, representing outcomes. For each unit i, we assume $\mathbf{L}_i \subseteq \mathrm{pa}_{\mathcal{G}}(A_i)$, and $\mathbf{L}_i \cup \{A_i\} \subseteq \mathrm{pa}_{\mathcal{G}}(Y_i)$. This represents a common assumption (which we call *causal ordering*) in causal inference that for each unit both baseline factors and treatment potentially affect the outcome, and that the baseline factors also affect treatment assignment. Here each unit has one treatment variable A_i, one outcome variable Y_i, and possibly many baseline variables $\mathbf{L}_i$. In interference settings, it is standard to allow that variables for another unit j may influence variables for unit i. In our case, there is a further complication: the precise nature of this influence is unknown.

This model implies, for positive $p(\mathbf{V})$, the following standard assumptions from the interference literature: $\mathbf{Y}(\mathbf{a}) \perp\!\!\!\perp \mathbf{A} \mid \mathbf{L}$ (network ignorability); $p(\mathbf{a} \mid \mathbf{L}) > 0$ $\forall \mathbf{a}$ (positivity); and $\mathbf{Y}(\mathbf{a}) = \mathbf{Y}$ if $\mathbf{A} = \mathbf{a}$ (consistency). Under these assumptions, the joint counterfactual outcome is identified, regardless of the underlying network structure, as the following special case of (2): $p(\mathbf{Y}(\mathbf{a})) = \sum_{\mathbf{L}} p(\mathbf{Y} \mid \mathbf{A} = \mathbf{a}, \mathbf{L}) p(\mathbf{L})$.

Given a particular treatment assignment probability $\pi(\mathbf{A})$, a number of causal effects of interest may be defined, see [36] for an extensive discussion. In this paper, we focus on a single effect, the *population average overall effect (PAOE)*, though our results generalize to any

identified causal effect of interest in network settings (for example, spillover effects). Consider a block is of size m and two fixed π_1, π_2 assignment probabilities. Then the PAOE is defined as:

$$\frac{1}{m} \sum_{i=1}^{m} \sum_{\mathbf{A}} \mathbb{E}[Y_i(\mathbf{A})]\{\pi_1(\mathbf{A}) - \pi_2(\mathbf{A})\}. \tag{3}$$

Under the aforementioned assumptions, this effect is identified by the following functional [36]:

$$\frac{1}{m} \sum_{i=1}^{m} \sum_{\mathbf{L}, \mathbf{A}} \mathbb{E}[Y_i \mid \mathbf{A}, \mathbf{L}] p(\mathbf{L})\{\pi_1(\mathbf{A}) - \pi_2(\mathbf{A})\}. \tag{4}$$

A number of estimation strategies for (4) are possible under various assumptions on network structure. For example, [36] considered an inverse probability weighted estimator. In this paper, we use the auto-g-computation algorithm in [35] to estimate the PAOE, which allows for arbitrary network structure; we describe this estimator in detail in the Supplement.

4 MODEL SELECTION FOR UNKNOWN NETWORKS

We are interested in estimating causal effects like the PAOE under the aforementioned assumptions, where there is uncertainty about the network structure. We give a taxonomy of problems of this type, having different levels of difficulty depending on the degree of uncertainty present.

The most general version of the problem occurs when neither the causal structure of each unit, nor the network structure inducing dependence between units, is known. In this case the problem reduces to a structure learning problem for arbitrary chain graphs, as considered in [18] and [22]. We do not pursue this version of the problem here for two reasons. First, the causal structure for each unit is often known due to background knowledge on temporal ordering and study design, as is the case for our needle-dispensary motivating example. Second, model selection of arbitrary CGs is known to be a very challenging problem which (in the worst case) may require large sample sizes [6].

In many settings, the causal structure for each individual unit is known and is typically assumed to be the same for every unit, i.e., $\mathcal{E}_i = \mathcal{E}_j$ for all i, j. The problem of model selection then amounts to learning the structure of the connections between units i.e., $\mathcal{E}_{ij}$ for all i, j. The search space for such a problem, while much smaller than the general problem, is still exponential. For a block that contains m units, there are $\binom{m}{2}$ possible pairings of

units, leading to $2^{\binom{m}{2}}$ possible networks. The number of possible valid chain graphs is even larger, since units i, j adjacent in a network could be connected in a variety of ways via (undirected or directed) edges in $\mathcal{E}_{ij}$. Learning these connections requires a search through all possible combinations of edges that form $\mathcal{E}_{ij}$ such that the overall graph is a CG.

We may restrict the problem further by requiring that the connections between any two units, if present, are *homogenous*, meaning that dependence between any two units, if it exists, arises in the same way. Formally, we define homogeneity such that, for all pairs $(i, j), (k, l) \in \mathcal{N}$, $\mathcal{E}_{ij} = \mathcal{E}_{kl}$. Notice that the space of homogenous networks is still fairly large. The problem may be made more tractable by one of the following two assumptions. We may assume the existence of network connections is known, but that their types are unknown, i.e., we know $\mathcal{N}$ and would like to learn $\mathcal{E}_{ij}$. Alternatively, we may assume we know how two adjacent units are connected, but not which pairs are adjacent, i.e., we know $\mathcal{E}_{ij}$ and would like to learn $\mathcal{N}$. We may also have no such background knowledge. In the following, we present algorithms for both homogenous and heterogenous settings.

Throughout, we make an assumption which we call *tier symmetry*, which is commonly made implicitly or explicitly in the interference literature [36, 35]. That is, we require connections between variables in the same "tier" of causal ordering to represent symmetric relations between the variables. This restricts edges $L_i - L_j$, $A_i - A_j$, and $Y_i - Y_j$ to always be undirected. Also it is natural to extend the known causal ordering of variables to connections between units: while we allow for e.g., $A_i \rightarrow Y_j$, the reverse, $Y_j \rightarrow A_i$ is ruled out. Finally, we rule out the existence of undirected edges connecting variables across tiers, e.g, edges of the form $A_i - Y_j$, since the existence of such edges, coupled with our causal ordering assumption, leads to graphs which are not CGs.

Before presenting algorithms to address the above taxonomy of problems, we introduce some necessary concepts from the graphical model selection literature.

4.1 Markov Properties and Faithfulness

If $p(\mathbf{V})$ is a positive distribution, the factorization (1) is equivalent to a global Markov property which relates certain graphical separation facts in the CG $\mathcal{G}$ (given by the c-separation criterion) to conditional independence relations in $p(\mathbf{V})$; see [14] for precise definitions. In what follows, we make the *faithfulness* assumption, which is the converse of the global Markov property: if $(\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C})$ in $p(\mathbf{V})$, then $\mathbf{A}$ is c-separated from $\mathbf{B}$ given $\mathbf{C}$ in $\mathcal{G}$. This is directly analogous to the faithfulness as-

sumption made when selecting DAG models from data by constraint-based or score-based methods [33, 4].

4.2 Model Scores and the Pseudolikelihood

In this paper, we will learn the structure of the network using a score-based approach to model selection. Score-based methods proceed by choosing the graph (from among some space of candidates) that optimizes a model score. Exhaustive model search is typically infeasible, so it is popular to employ greedy methods that optimize only "locally," that is, they traverse the space of candidate graphs considering only single-edge additions and deletions. Under some conditions, such greedy procedures can be shown to asymptotically converge to the globally optimal model [4]. Scores used for greedy search typically satisfy three properties that are sufficient for finding the globally optimal model: decomposability, score-equivalence, and consistency.

A score is said to be decomposable if it can be written as a sum of local contributions, each a function of one vertex and its boundary. A score is said to be score-equivalent if two Markov equivalent graphs (i.e., graphs that imply the same set of conditional independences by the global Markov property) yield the same score. A score is said to be consistent if, as the sample size goes to infinity, the following two conditions hold. First, when two models both contain the true generating model, the model of lower dimension will have a better score. Second, when one model contains the true model and another does not, the former will have a better score.

A popular score satisfying these properties for model selection among DAG models is the Bayesian Information Criterion (BIC) [28]. Given a d-dimensional data set $\mathbf{D}$ of size n and model likelihood $\mathcal{L}(\mathbf{D}; \mathcal{G}) \equiv \prod_{i=1}^{n} p(x_{1,i}, \ldots, x_{d,i}; \mathcal{G})$, the BIC is given by $2 \ln \mathcal{L}(\mathbf{D}; \mathcal{G}) - k \ln(n)$ where k is model dimension.

For CG models, the BIC is only decomposable for blocks, not for variables within the block. In addition, the score is not easy to evaluate. Both of these issues arise due to the presence of normalizing functions in the likelihood. Here, we present an alternative score which avoids some of these problems, based on the *pseudolikelihood function* [2]:

$$\mathcal{PL}(\mathbf{D}; \mathcal{G}) \equiv \prod_{i=1}^{n} \prod_{j=1}^{d} p(x_{j,i} \mid \mathbf{x}_{-j,i}; \mathcal{G}),$$

where $\mathbf{x}_{-j}$ is the vector $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d)$. We define a score based on the pseudolikelihood called Pseudo-BIC (PBIC): $2 \ln \mathcal{PL}(\mathbf{D}; \mathcal{G}) - k \ln(n)$.

We propose a greedy score-based model selection proce-

dure based on the PBIC score, which is consistent and obeys a weaker notion of decomposability for exponential families, as we show below. All proofs are deferred to the Supplement.

Lemma 1. *With dimension fixed and sample size increasing to infinity, the PBIC is a consistent score for curved exponential families whose natural parameter space Θ forms a compact set.*

Decomposability of a scoring criterion makes greedy search a practical procedure, by limiting the number of terms in the overall score that need to be recomputed for each considered edge modification. While the BIC score for DAG models is decomposable, the PBIC score for CG models is not. Nevertheless, a weaker notion of decomposability holds, which implies that two CG models that differ by a single edge differ by a subset of components of the score, which we now describe. Consider a candidate edge between V_i and V_j in a CG $\mathcal{G}$. Let $\mathbf{B}_{\text{loc}}$ denote the block to which V_j belongs when the edge is directed $V_i \to V_j$, or to which V_i and V_j belong when the edge is undirected $V_i - V_j$. We use $\text{loc}(V_i, V_j; \mathcal{G})$ to denote a set of vertices called the *local set*, defined as:

$$\bigcup_{\mathbf{C}} \{\mathbf{C} \in \mathcal{C}((\mathcal{G}_{\text{bd}_{\mathcal{G}}(\mathbf{B}_{\text{loc}})})^a) : V_i, V_j \in \mathbf{C} \not\subseteq \text{pa}_{\mathcal{G}}(\mathbf{B}_{\text{loc}})\}.$$

As we show, the score difference for graphs $\mathcal{G}$ and $\mathcal{G}'$ which differ by a single edge can be written as the difference between terms that involve only variables in the local set of $\mathcal{G}$. The next result, and much subsequent discussion in the paper, is stated for conditional Markov random fields (MRFs). This is because statistical CG models can be equivalently described as sets of conditional MRF models. We elaborate on this relationship in the Supplement.

Lemma 2. *Let $\mathcal{G}$ and $\mathcal{G}'$ be graphs which differ by a single edge between V_i and V_j. For conditional MRFs in the exponential family, the local score difference between $\mathcal{G}$ and $\mathcal{G}'$ is given by:* $\sum_{V \in \text{loc}(V_i, V_j; \mathcal{G}) \cap \mathbf{B}_{\text{loc}}} \{s_V(\mathbf{D}; \mathcal{G}) - s_V(\mathbf{D}; \mathcal{G}')\}$, *where $s_V(.)$ denotes the component of the score for V.*

Note that the above definition of the local set may simplify further in certain special cases of MRF models in the exponential family. In particular, if we consider an MRF that is multivariate normal, or a log linear discrete model with only main effects and pairwise interactions, then the sum in Lemma 2 reduces to either a sum over elements V_i and V_j (for an undirected edge $V_i - V_j$) or only V_j (for a directed edge $V_i \to V_j$). We omit the straightforward proofs in the interest of space. We will not consider these special instances of the exponential family in the remainder of this paper, but in the supplement we

---

Algorithm 1 GREEDY NETWORK SEARCH($\mathcal{G}^{\text{init}}, \mathbf{D}$)

1: $\mathcal{G}^* \leftarrow \mathcal{G}^{\text{init}}$
2: score change $\leftarrow$ True
3: **while** score change **do**
4:      score change $\leftarrow$ False
5:      $\mathcal{E}_{\mathcal{N}}^* \leftarrow$ network ties in $\mathcal{G}^*$
6:      $E_{max} \leftarrow \text{argmax}_{E \in \mathcal{E}_{\mathcal{N}}^*} \text{PBIC}(\mathbf{D}; \mathcal{G}^* \setminus E)$
7:      **if** $\text{PBIC}(\mathbf{D}; \mathcal{G}^* \setminus E_{max}) > \text{PBIC}(\mathbf{D}; \mathcal{G}^*)$ **then**
8:          $\mathcal{G}^* \leftarrow \mathcal{G}^* \setminus E_{max}$ $\triangleright$ deleting edge E_{max}
9:          score change $\leftarrow$ True
10: **return** $\mathcal{E}_{\mathcal{N}}^*$

---

discuss the incurred computational costs for exponential families in general.

4.3 Greedy Network Search

While there exist numerous methods that take a pseudolikelihood-type approach to model selection in UGs [24, 11, 7, 1], these have been typically restricted to Ising or Gaussian models. Such methods involve a per-vertex neighbourhood selection procedure using L1-regularized regression or the standard BIC, which may yield self-inconsistent results (e.g., find that V_i in $\text{nb}(V_j)$ but not vice versa). Any resulting inconsistencies would need to be resolved post hoc through union or intersection consolidation procedures. Methods that try to enforce self-consistency by explicitly maximizing the pseudolikelihood with a regularization penalty are presented in [8] and [12], but are again restricted to Ising and Gaussian graphical models. The properties of the PBIC described in the previous section allow us to design algorithms for greedy network search that are parallelizable, while also generalizing to all exponential families and circumventing the need for post hoc procedures. While our method covers a more general class of models, it can be computationally expensive to calculate the local scores at each step. A more efficient procedure is possible in some subclasses (including Ising and Gaussian), where we can modify our procedure into a "forward-backward" algorithm reminiscent of the GES algorithm [4]. Since our focus is on a general procedure for all exponential families, we defer further discussion of these special cases to the Supplement.

We begin by describing a greedy search procedure that learns network ties $\mathcal{E}_{\mathcal{N}}$, without imposing homogeneity. Model selection proceeds by solving 3 independent sub problems: learning a Markov random field (MRF) over the baseline covariates $\mathbf{L}$, learning a conditional MRF on the treatments $\mathbf{A}$, and learning a conditional MRF on the outcomes $\mathbf{Y}$. The resulting network ties learned from each of these, are combined to produce the final re-

---

Algorithm 2 HETEROGENOUS($\mathcal{G}^{\text{complete}}, \mathbf{D}$)

1: $\mathcal{G}^{\mathbf{L}}, \mathcal{G}^{\mathbf{A}}, \mathcal{G}^{\mathbf{Y}} \leftarrow$ conditional MRFs on $\mathbf{L}$, $\mathbf{A}$, and $\mathbf{Y}$ formed from $\mathcal{G}^{\text{complete}}$
2: $\mathcal{E}^*_{\mathcal{N}_{\mathbf{L}}} \leftarrow$ GREEDY NETWORK SEARCH($\mathcal{G}^{\mathbf{L}}, \mathbf{D}$)
3: $\mathcal{E}^*_{\mathcal{N}_{\mathbf{A}}} \leftarrow$ GREEDY NETWORK SEARCH($\mathcal{G}^{\mathbf{A}}, \mathbf{D}$)
4: $\mathcal{E}^*_{\mathcal{N}_{\mathbf{Y}}} \leftarrow$ GREEDY NETWORK SEARCH($\mathcal{G}^{\mathbf{Y}}, \mathbf{D}$)
5: **return** $\mathcal{E}^*_{\mathcal{N}_{\mathbf{L}}} \cup \mathcal{E}^*_{\mathcal{N}_{\mathbf{A}}} \cup \mathcal{E}^*_{\mathcal{N}_{\mathbf{Y}}}$

Algorithm 3 HOMOGENOUS($\mathcal{G}^{\text{complete}}, \mathbf{D}, \mathcal{E}_{\text{proto}\,\mathcal{N}}$)

1: $\mathcal{G}^* \leftarrow$ graph obtained by removing all edges between units i, j in $\mathcal{G}^{\text{complete}}$ when $E_{ij} \notin h(\mathcal{E}_{\text{proto}\,\mathcal{N}})$
2: score change $\leftarrow$ True
3: **while** score change **do**
4:     score change $\leftarrow$ False
5:     $\mathcal{N}^* \leftarrow$ network in $\mathcal{G}^*$
6:     $(i,j)_{max} \leftarrow \arg\max_{(i,j)\in\mathcal{N}^*} \text{PBIC}(\mathbf{D}; \mathcal{G}^* \setminus \mathcal{E}_{ij})$
7:     **if** $\text{PBIC}(\mathbf{D}; \mathcal{G}^* \setminus \mathcal{E}_{ij_{max}}) > \text{PBIC}(\mathbf{D}; \mathcal{G}^*)$ **then**
8:         $\mathcal{G}^* \leftarrow \mathcal{G}^* \setminus \mathcal{E}_{ij_{max}}$
9:         score change $\leftarrow$ True
10: **return** $\mathcal{N}^*$

Algorithm 4 HOMOGENOUS($\mathcal{G}^{\text{complete}}, \mathbf{D}, \mathcal{N}$)

1: $\mathcal{G}^* \leftarrow$ graph obtained by removing all edges between units i, j in $\mathcal{G}^{\text{complete}}$ when $(i,j) \notin \mathcal{N}$
2: score change $\leftarrow$ True
3: **while** score change **do**
4:     score change $\leftarrow$ False
5:     $\mathcal{E}^*_{\text{proto}\,\mathcal{N}} \leftarrow$ prototypes of network ties in $\mathcal{G}^*$
6:     $E_{max} \leftarrow \arg\max_{E\in\mathcal{E}^*_{\text{proto}\,\mathcal{N}}} \text{PBIC}(\mathbf{D}; \mathcal{G}^* \setminus h(E))$
7:     **if** $\text{PBIC}(\mathbf{D}; \mathcal{G}^* \setminus h(E_{max})) > \text{PBIC}(\mathbf{D}; \mathcal{G}^*)$ **then**
8:         $\mathcal{G}^* \leftarrow \mathcal{G}^* \setminus h(E_{max})$
9:         score change $\leftarrow$ True
10: **return** $\mathcal{E}^*_{\text{proto}\,\mathcal{N}}$

Algorithm 5 HOMOGENOUS($\mathcal{G}^{\text{complete}}, \mathbf{D}$)

1: $\mathcal{E}_{\text{proto}\,\mathcal{N}} \leftarrow$ prototypes of network ties in $\mathcal{G}^{\text{complete}}$
2: $\mathcal{N}^* \leftarrow$ HOMOGENOUS($\mathcal{G}^{\text{complete}}, \mathbf{D}, \mathcal{E}_{\text{proto}\,\mathcal{N}}$)
3: $\mathcal{E}^*_{\text{proto}\,\mathcal{N}} \leftarrow$ HOMOGENOUS($\mathcal{G}^{\text{complete}}, \mathbf{D}, \mathcal{N}^*$)
4: **return** $\mathcal{N}^*, \mathcal{E}^*_{\text{proto}\,\mathcal{N}}$

---

sult (Alg. 2). Each of the above subproblems is solved by a greedy search procedure (Alg. 1) that starts with the complete conditional MRF (or MRF), and deletes the edge that yields the greatest improvement to the PBIC score on each iteration.

We now describe procedures for learning network ties in the homogenous setting, after defining some preliminaries. The *homologs* of an edge $E_{ij} \in \mathcal{E}_{\mathcal{N}}$ with endpoints $U_i, W_j \in \mathbf{V}$, are defined as: $h(E_{ij}) \equiv \{E_{kl} \in \mathcal{E}_{\mathcal{N}} : \text{endpoints}(E_{kl}) = U_k, W_l\}$. The network tie prototypes in a homogenous graph $\mathcal{G}$ are defined as: $\mathcal{E}_{\text{proto}\,\mathcal{N}} \equiv \{E_{ij} \in \mathcal{E}_{ij} \text{ for any } (i,j) \in \mathcal{N}\}$. $h(\mathcal{E}_{\text{proto}\,\mathcal{N}})$ can then be defined as: $\{h(E) : E \in \mathcal{E}_{\text{proto}\,\mathcal{N}}\}$.

When the types of connections $\mathcal{E}_{\text{proto}\,\mathcal{N}}$ between any two connected units is known, we start with a CG that is fully connected as $\mathcal{E}_{\text{proto}\,\mathcal{N}}$ for every pairwise combination of units. Search proceeds by deleting $\mathcal{E}_{ij}$ between two units i and j that yields the best improvement in the PBIC on each iteration (Alg. 3). When the social network $\mathcal{N}$ is known, we start with a CG where pairs of units in $\mathcal{N}$ are fully connected in network ties. Search proceeds by deleting all homologs of the type of edge in $\mathcal{E}_{\text{proto}\,\mathcal{N}}$ that yields the best improvement in the PBIC on each iteration (Alg. 4). Finally, when there is no background knowledge, homogenous search (Alg. 5) can be performed by chaining the operations of Alg. 3 and Alg. 4 (or vice versa) on the CG complete in network ties for every pairwise combination of units.

Clearly we could use the heterogenous procedure even if the true underlying network ties are homogenous, since it is most general. However, intuitively we expect the homogenous procedures to fare better in a finite data setting, because the homogeneity assumption allows pooling data from samples across units for each edge deletion test. This intuition is confirmed in our simulations.

4.4 Size of the Search Space

In the heterogenous case, the search space grows as $O(|\mathcal{E}_{\text{proto}\,\mathcal{N}}|\binom{m}{2})$ i.e., as a function of the number of possible edges between two units i and j multiplied by the number of possible pairings on m units. Under homogeneity when $\mathcal{E}_{\text{proto}\,\mathcal{N}}$ is known, this reduces to $O(\binom{m}{2})$; when $\mathcal{N}$ is known, it reduces to $O(|\mathcal{E}_{\text{proto}\,\mathcal{N}}|)$; and under homogeneity where neither is available, it is $O(\binom{m}{2}) + O(|\mathcal{E}_{\text{proto}\,\mathcal{N}}|)$.

4.5 Consistency of Network Search

Lemma 3. *If the generating distribution is Markov to a CG satisfying tier symmetry and the causal ordering assumption, then the search space of* GREEDY NETWORK SEARCH *consists of graphs belonging to their own equivalence classes of size 1.*

Theorem 1. *If the generating distribution is in the exponential family (with compact natural parameter space Θ) and is Markov and faithful to a CG satisfying tier symmetry and causal ordering, then* GREEDY NETWORK SEARCH *is consistent.*

Under the same assumptions in the theorem above, we

have the following corollary results.

Corollary 1.1. *The* HETEROGENOUS *procedure is consistent.*

Corollary 1.2. *When the true network ties are homogenous, the* HOMOGENOUS *procedure is consistent.*

5 EXPERIMENTS

We evaluate the performance of our proposed algorithms on networks of varying size, for various block sizes, and for different regularity settings. (Regularity refers to the number of neighbors for each unit i in the dependency network $\mathcal{N}$. This setting thus controls the density of the graph.) We consider blocks of size 4, 8, 16, and 32, with regularity 2 or 3. The ground truth models are homogenous and of the form shown in Figures 2 and 3, where we display the case of block size 4. Data is generated from each network via a Gibbs sampler with a burn-in period of 1000 iterations and thinning every 100 iterations using the following equations:

$$p(L_i = 1) = \text{expit}(\tau_1),$$
$$p(A_i = 1 | L_i, \{A_j : j \in \text{nb}_\mathcal{N}(i)\}) =$$
$$\text{expit}(\beta_1 L_i + \beta_2 \sum_{j \in \text{nb}_\mathcal{N}(i)} A_j),$$
$$p(Y_i = 1 | L_i, A_i, \{A_j : j \in \text{nb}_\mathcal{N}(i)\}) =$$
$$\text{expit}(\nu_1 L_i + \nu_2 A_i + \nu_3 \sum_{j \in \text{nb}_\mathcal{N}(i)} A_j),$$

where $\text{expit}(x) = (1 + \exp(-x))^{-1}$. We emphasize that some of these networks are quite large; for example, the network with block size 32 and 2000 iid blocks has an effective size of 64,000 individuals. For each network setting we run 100 bootstraps of structure learning in order to get an average estimate of precision and recall as shown in Figure 4. However, to spare computation time, we use only Algorithm 3 on the latter two block settings. An interesting feature of the results in Figure 4, which matches our earlier intuition, is the faster convergence of the homogenous procedures to the true model – which we attribute to the parameter sharing (effectively using of more data when testing each edge deletion).

In order to demonstrate the utility of learning the structure in dealing with network uncertainty, we consider the population average overall effect (3). We first execute structure learning, and then estimate the PAOE, contrasting a treatment assignment determined with probability 0.7 with the naturally observed probability. We do this for 2-regular networks with 2000 realizations of iid blocks of varying size. We use the heterogenous procedure and one of the homogenous procedures (Alg. 3) to learn the structure of the networks. Estimation of the

causal effect is done by the auto-g-computation algorithm described in [35] and the Supplement. We perform a 1000 bootstraps of both structure learning and effect estimation to compare the bias and variance of the estimates from the learned graphs to the estimates provided by utilizing the maximally uninformative complete graph. Unfortunately the auto-g-computation procedure is also computationally intensive because it requires Gibbs sampling. Again, to spare computation time we do not run the heterogenous procedure on the larger graphs with block sizes 16 and 32 (networks with 32,000 and 64,000 individuals). We also only perform 8 bootstraps for these larger networks. In order to emphasize the need to deal with interference and network uncertainty appropriately, we additionally estimated the bias for 200 bootstraps of the network with blocks of size 8 using the empty graph (a complete iid assumption), and an incorrect graph where $\mathcal{N}$ is shuffled randomly to have incorrect adjacencies. In both cases the bias turned out to be approximately .06, an order of magnitude higher than the bias from utilizing the complete or learned graphs.
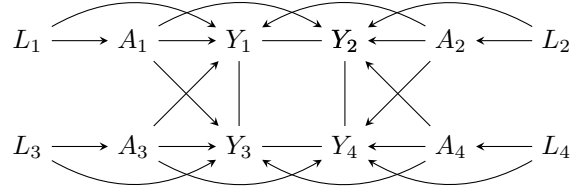

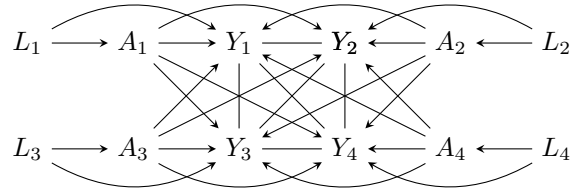Figure 2: The 2-regular CG for a block of size 4

Figure 3: The 3-regular CG for a block of size 4

From Table 1 we see that causal effect estimates based on learned structure have the same or lower bias as com-

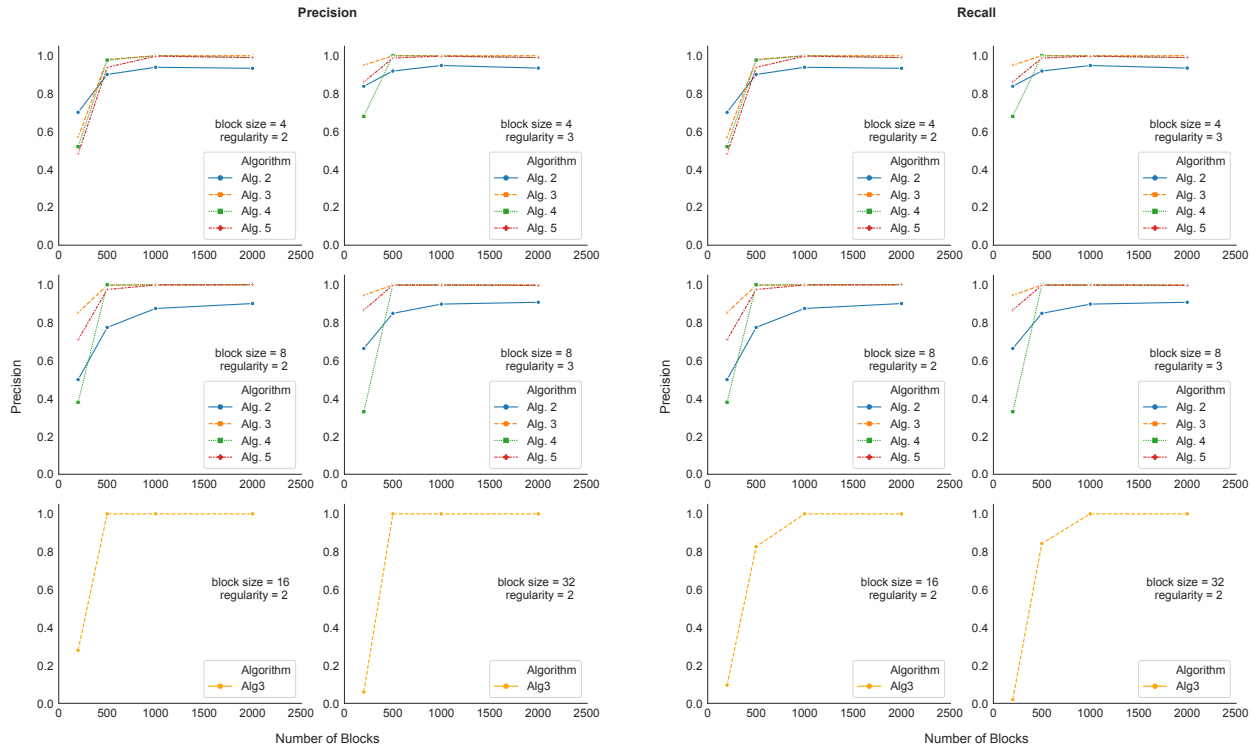| Block Size	Complete	Homogenous	Heterogenous
4	.009, 9.2e-5	.008, 8.1e-5	.009, 9.7e-5
8	.007, 6.6e-5	.006, 4.1e-5	.006, 4.5e-5
16	.006, 3.8e-5	.005, 1.9e-5	x
32	.007, 6.1e-5	.002, 7.6e-6	x

Table 1: Bias and variance for estimating the PAOE.

Figure 4: Performance of structure learning algorithms as measured by precision and recall

pared with using the complete graph. Furthermore, the sparsity of the learned graph reduces variance of the estimates in most cases. This reduction in bias and variance is more easily achieved when we are able to exploit homogeneity in the network structure. In experiments with lower sample sizes, we see that the bias of effect estimates may increase (because the learning procedure may fail to recover the true graph) but that the variance of the estimates remains comparable to or lower than the estimates based on the complete graph.

6 CONCLUSION

We have developed a method for estimating causal effects under unit dependence induced by a network represented by a chain graph (CG) model [14], when there is uncertainty about network structure. Instead of estimating causal effects given a completely uninformative network where each pair of units is connected, as is typically done in the interference literature [36, 37], we estimated causal effects given a sparser network learned via a score-based model selection method based on the pseudolikelihood function [2]. We showed that this strategy can yield lower variance in estimates without sacrificing bias, if the underlying true network structure is recovered accurately. Our model selection method re-

lied on weak parametric assumptions, specifically that all Markov factors in the CG model corresponded to conditional Markov random fields in the exponential family. The approach here is a generalization of local score-based search algorithms for directed acyclic graph (DAG) models [4] to CG models. As a price of this generalization, our local search algorithms recompute a potentially larger part of the model score with every move through the model space. In addition, our approach only works for settings with *partial interference*, where units within a block exhibit dependence, but data on blocks is iid. The restriction to blocks of identical size may be relaxed by combining our heterogeneous procedure with a scheme of parameter sharing and hierarchical modeling across blocks that are of different sizes. In future work, we aim to extend our methods to *full interference* settings.

Acknowledgements

References

[1] Rina Foygel Barber and Mathias Drton. High-dimensional Ising model selection with Bayesian information criteria. *Electronic Journal of Statistics*, 9(1):567–607, 2015.

[2] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological Statistics)*, 36(2):192–236, 1974.

[3] Yann Bramoullé, Andrea Galeotti, and Brian Rogers. *The Oxford Handbook of the Economics of Networks*. Oxford University Press, 2016.

[4] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

[5] Forrest W. Crawford, Peter M. Aronow, Li Zeng, and Jianghong Li. Identification of homophily and preferential recruitment in respondent-driven sampling. *American Journal of Epidemiology*, 187(1):153–160, 2017.

[6] Robin J. Evans. Model selection and local geometry. *arXiv preprint arXiv:1801.08364*, 2018.

[7] Rina Foygel and Mathias Drton. Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems*, pages 604–612, 2010.

[8] Holger Höfling and Robert Tibshirani. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*, 10(Apr):883–906, 2009.

[9] Guanglei Hong and Stephen W. Raudenbush. Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101(475):901–910, 2006.

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

[11] Ali Jalali, Christopher C. Johnson, and Pradeep K. Ravikumar. On learning discrete graphical models using greedy methods. In *Advances in Neural Information Processing Systems*, pages 1935–1943, 2011.

[12] Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):803–825, 2015.

[13] Adam D. I. Kramer, Jamie E Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, pages 8788–8790, 2014.

[14] Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[15] Steffen L. Lauritzen and Thomas S. Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):321–348, 2002.

[16] Sanghack Lee and Vasant Honavar. On learning causal models from relational data. In *Thirtieth AAAI Conference on Artificial Intelligence*, pages 3263–3270, 2016.

[17] Kevin Lewis, Marco Gonzalez, and Jason Kaufman. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109(1):68–72, 2012.

[18] Zongming Ma, Xianchao Xie, and Zhi Geng. Structural learning of chain graphs via decomposition. *Journal of Machine Learning Research*, 9(Dec):2847–2880, 2008.

[19] Marc Maier, Katerina Marazopoulou, David Arbour, and David Jensen. A sound and complete algorithm for learning causal models from relational data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 371–380. AUAI Press, 2013.

[20] Elizabeth L. Ogburn and Tyler J. VanderWeele. Causal diagrams for interference. *Statistical Science*, 29(4):559–578, 2014.

[21] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[22] Jose Peña, Dag Sonntag, and Jens Nielsen. An inclusion optimal algorithm for chain graph structure learning. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 778–786, 2014.

[23] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

[24] Pradeep K. Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using L1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

[25] Thomas S. Richardson and James M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper 128*, pages 1–146, 2013.

[26] James M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

[27] Paul R. Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200, 2007.

[28] Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[29] 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, 2011.

[30] Eli Sherman and Ilya Shpitser. Identification and estimation of causal effects from dependent data. In *Advances in Neural Information Processing Systems 31*, pages 9424–9435. 2018.

[31] Shohei Shimizu. LiNGAM: non-Gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

[32] Michael E. Sobel. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.

[33] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

[34] Sharon Stancliff, Bruce Agins, Josiah D. Rich, and Scott Burris. Syringe access for the prevention of blood borne infections among injection drug users. *BMC Public Health*, 3(1):37, 2003.

[35] Eric J. Tchetgen Tchetgen, Isabel Fulcher, and Ilya Shpitser. Auto-G-Computation of causal effects on a network. *arXiv:1709.01577*, 2017.

[36] Eric J. Tchetgen Tchetgen and Tyler J. VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.

[37] Tyler J. VanderWeele, Eric J. Tchetgen Tchetgen, and M. Elizabeth Halloran. Components of the indirect effect in vaccine trials: identification of contagion and infectiousness effects. *Epidemiology*, 23(5):751, 2012.