Supplement For: 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

CAUSAL CHAIN GRAPHS AND THEIR INTERPRETATION

Causal models associated with DAGs may be generalized to causal models associated with CGs. CGs may include directed edges, representing direct causation, and undirected edges, representing symmetric relationships between units in a network. A causal interpretation of CGs, understood as equilibria of dynamic models with feedback, was given in [5]. Under this interpretation, the distribution $p(\mathbf{B} \mid \mathrm{pa}_{\mathcal{G}}(\mathbf{B}))$ for each block $\mathbf{B} \in \mathcal{B}(\mathcal{G})$ can be determined by a Gibbs sampler on the variables $B \in \mathbf{B}$. Here, each conditional distribution $p(B \mid \mathbf{B} \setminus B, \mathrm{pa}_{\mathcal{G}}(\mathbf{B}))$ is produced by structural equations of the form $f_B(\mathbf{B} \setminus B, \mathrm{pa}_{\mathcal{G}}(\mathbf{B}), \epsilon_B)$. Interventions on elements of $\mathbf{B}$ are defined by replacing the appropriate line in the Gibbs sampler program. For all disjoint sets $\mathbf{Y}$ and $\mathbf{A}$, [5] showed that $p(\mathbf{Y} \mid \mathrm{do}(\mathbf{a}))$ is identified by a CG version of the g-formula (2).

If only interventions on entire blocks are of interest, i.e., we consider only treatment assignments $\mathbf{A}$ such that if $\mathbf{B} \cap \mathbf{A} \neq \emptyset$ then $\mathbf{B} \subseteq \mathbf{A}$, then an alternative causal interpretation of a CG $\mathcal{G}$ that does not rely on the Gibbs sampler machinery of [5] exists. Specifically, in such a case we consider a causal DAG model where each block $\mathbf{B}$ corresponds to a supervariable $V_{\mathbf{B}}$ defined as a Cartesian product of variables in $\mathbf{B}$, and a DAG causal model is defined on $V_{\mathbf{B}}(\mathbf{A})$, where $\mathbf{A}$ are values assigned to parents of $V_{\mathbf{B}}$.

If, for each block $\mathbf{B}$ in a CG $\mathcal{G}$, the graph $(\mathcal{G}_{\mathrm{bd}_{\mathcal{G}}(\mathbf{B})})^a$ has a single clique, then this yields a classical causal model of a DAG, defined on $\{V_{\mathbf{B}} | \mathbf{B} \in \mathcal{B}(\mathcal{G})\}$. If not, we can still view the model as a classical causal model of a DAG, but with an extra restriction that the observed data distribution factorizes as (1). See also [7] for a perspective on interpreting chain graphs in an interference setting.

The model selection methodology introduced here does not depend on which causal interpretation for chain graphs one may choose, and all causal models described above lead to interventional distributions being identified by (2).

CONDITIONAL MRFs

A CG model can be viewed as a set of conditional MRFs. A conditional MRF corresponds to a graph whose vertices can be partitioned into two disjoint sets: $\mathbf{W}$, corresponding to non-random variables whose values are fixed; and $\mathbf{V}$, corresponding to random variables. The only edges allowed in a conditional MRF are directed edges $W \rightarrow V$ and undirected edges $V - V'$
for $W \in \mathbf{W}$ and $V, V' \in \mathbf{V}$. A statistical model associated with a conditional MRF $\mathcal{G}$ is a set of densities that factorize as:

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

It is easy to see that the above factorization is analogous to the second level of CG factorization found in (1) where $\mathbf{V}$ is a block, and $\mathbf{W}$ are its parents.

THE AUTO-G-COMPUTATION ALGORITHM

The auto-g-computation algorithm, introduced in [8], may be viewed as a generalization of the Monte Carlo sampling version of the g-computation algorithm for classical causal models (represented by DAGs) [9] to causal models of the sort we consider here, represented by CGs. We describe a version of this algorithm based on the pseudolikelihood estimator. An alternative based on the coding estimator [1] is less efficient, but leads to asymptotically normal estimators of the population average overall effect (PAOE).

Auto-g-computation generates samples from either the observed data distribution that factorizes as (1) according

to a CG, or of functions of these distributions, such as counterfactual expectations identified using (4).

This is done by imposing a topological ordering on blocks in a CG, and generating samples for each block sequentially using Gibbs sampling. The parameters for Gibbs factors used in the sampler (which by the global Markov property for CGs take the form of $p(X_i|X_{\mathrm{bd}_\mathcal{G}(X_i)})$) are learned via maximizing the pseudolikelihood function. For any block $\mathbf{X}$, the Gibbs sampler draws samples from $p(\mathbf{X} \mid \mathrm{bd}_\mathcal{G}(\mathbf{X}))$, given a fixed set of samples drawn from all blocks with elements in $\mathrm{pa}_\mathcal{G}(\mathbf{X})$ as follows:

Gibbs Sampler for $\mathbf{X}$:

for $t = 0$, let $\mathbf{x}^{(0)}$ denote initial values ;
for $t = 1, ..., T$

draw value of $X_1^{(t)}$ from $p(X_1|\mathbf{x}^{(t-1)}_{\mathrm{bd}_\mathcal{G}(X_1)})$);

draw value of $X_2^{(t)}$ from $p(X_2|\mathbf{x}^{(t-1)}_{\mathrm{bd}_\mathcal{G}(X_2)})$);

$\vdots$

draw value of $X_m^{(t)}$ from $p(X_m|\mathbf{x}^{(t-1)}_{\mathrm{bd}_\mathcal{G}(X_m)})$);

This method may be used to estimate the counterfactual expectation in (4) as follows. We first generate a set of samples $\mathbf{L}^{(t)}$, $t = 1, \ldots, T$. Then we generate a sample $\mathbf{A}$ directly using some $\pi_i(\mathbf{A})$, $i = 1, 2$. Finally, we use the above samples to generate a set of samples $\mathbf{Y}^{(t)}$, $t = 1, \ldots, T$ using Gibbs factors $p(Y_i \mid \mathbf{A}_{\mathbf{A}\cap\mathrm{bd}_\mathcal{G}(Y_i)}, \mathrm{bd}_\mathcal{G}(Y_i) \setminus \mathbf{A})$. Finally, we estimate

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[Y_i(\mathbf{A})] = \frac{1}{m\cdot T}\sum_{i=1}^{m}\sum_{t=1}^{T} Y_i^{(t)}.$$

It is not difficult to show, (see [8] for details), that rerunning this procedure with different draws $\mathbf{A}$ from either $\pi_1(\mathbf{A})$ or $\pi_2(\mathbf{A})$, and taking the difference of the resulting averages yields a valid estimate of the PAOE.

Fitting parameters of Gibbs factors using the pseudolikelihood function avoids the usual difficulties CGs inherit from Markov random fields, specifically, the intractability of the likelihood function due to the presence of normalizing functions. In addition, if the learned block structure is sparse, while the number of independent samples considered is small, this approach allows one to impose parameter sharing among Gibbs factors, which leads to reasonable estimates even in small samples. Taken to the extreme, this approach allows inferences to be made even from a *single sample* of a network, as discussed in detail in [8]. In this manuscript we only consider the setting where multiple independent samples from blocks are available.

COMPUTATIONAL COMPLEXITY OF COMPUTING SCORES OF A CHAIN GRAPH MODEL

In blocks of a CG, the number of local terms that need to be computed corresponds to the number of vertices present in cliques containing the edge of interest in the augmented subgraph of the block and its parents. A term for V_j requires an $O(|\,\mathrm{bd}_\mathcal{G}(V_j)|)$ computation to update, which in the worst case may be exponential in the number of vertices if the graph is not sparse. In search problems, restrictions can be made on the maximum size of the boundary set, sacrificing accuracy for tractability. For a block in a CG corresponding to a conditional MRF in the exponential family, and an edge that is present in a set of cliques spanning all vertices, we will have a local set of size $O(d)$ in the worst case, with each local term requiring an $O(\text{clique size})$ computation. Thus, limiting the maximum clique size may speed up the computation of each local term, but in many cases we may be unable to avoid an $O(d)$ number of such terms. In other words, our scoring method for CG models where blocks correspond to conditional MRFs in the exponential family may not scale to very large graphs, even if such graphs are sparse. Achieving such a scaling will entail making additional assumptions, such as Gaussianity, or nonexistence of higher order interaction terms in log-linear models. We contrast this with DAG models, where the local set is of constant size regardless of parametric assumptions made.

FORWARD-BACKWARD SEARCH

Consistency of the score was sufficient to show consistency of a backwards greedy search involving only edge deletions starting from a complete conditional MRF. [2] showed that a property called *local consistency*, which follows from decomposability and consistency of the score, is sufficient to design a consistent forward-backward greedy search in the space of (Markov equivalent) DAGs. The forward stepwise search considers additions, rather than deletions, of single edges to improve the score, which typically produces a more sparse starting model for the subsequent backwards search.

Consider a graph $\mathcal{G}$ and another $\mathcal{G}'$ that differs only by the addition of an edge $V_i - V_j$ or $V_i \to V_j$. A score $S(\mathbf{D}; \mathcal{G})$ is called locally consistent if:

1. $V_i \not\perp\!\!\!\perp_{\mathcal{G}_0} V_j \mid \mathrm{bd}_\mathcal{G}(V_i)$ **or** $V_j \not\perp\!\!\!\perp_{\mathcal{G}_0} V_i \mid \mathrm{bd}_\mathcal{G}(V_j)$ then $\lim_{n\to\infty} P(S(\mathbf{D};\mathcal{G}') > S(\mathbf{D};\mathcal{G})) \to 1$
2. $V_i \perp\!\!\!\perp_{\mathcal{G}_0} V_j \mid \mathrm{bd}_{\mathcal{G}'}(V_i)$ **and** $V_j \perp\!\!\!\perp_{\mathcal{G}_0} V_i \mid \mathrm{bd}_\mathcal{G}(V_j)$ then $\lim_{n\to\infty} P(S(\mathbf{D};\mathcal{G}') < S(\mathbf{D};\mathcal{G})) \to 1$

Such a property requires a stronger notion of decomposability than is available in our general setting. In Section 4.2 we mention that if our model is an MRF that is multivariate normal, or corresponds to a log linear discrete model with only main effects and pairwise interactions, then it suffices to consider the following terms derived from the local set: $\{s(V_i, \mathrm{bd}_{\mathcal{G}}(V_i)), s(V_j, \mathrm{bd}_{\mathcal{G}}(V_j))\}$ for an edge $V_i - V_j$, and $\{s(V_j, \mathrm{bd}_{\mathcal{G}}(V_j))\}$ for an edge $V_i \rightarrow V_j$ (dropping implicit $\mathbf{D}$ and $\mathcal{G}$ for brevity). This is the strong notion of decomposability we need for local consistency. Thus, in such settings one can follow the work in [2] to show that PBIC will be locally consistent and design a search procedure involving a forward phase followed by a backward phase. The advantage of such a procedure is that it is more scalable, even more so when the underlying true model is sparse.

PROOFS

Let $\mathcal{M}_0$ denote the true model and $\mathcal{M}_1$, $\mathcal{M}_2$ two candidate models. A scoring criterion $S(\mathbf{D}; \mathcal{M})$ is said to be *consistent* if:

$$\lim_{n\to\infty} P_n(S(\mathbf{D};\mathcal{M}_1) < S(\mathbf{D};\mathcal{M}_2)) \rightarrow 1 \text{ when}$$
$$\mathcal{M}_1 \not\supseteq \mathcal{M}_0 \text{ and } \mathcal{M}_2 \supseteq \mathcal{M}_0 \text{ or} \quad (*)$$
$$\mathcal{M}_1, \mathcal{M}_2 \supseteq \mathcal{M}_0 \text{ and } k_1 > k_2. \quad (**)$$

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.*

Proof. To prove consistency we need to show that,

$$\lim_{n\to\infty} P_n(PBIC(\mathbf{D};\mathcal{M}_1) < PBIC(\mathbf{D};\mathcal{M}_2)) \rightarrow 1 \quad (1)$$

when (*) or (**).

Note in all following steps, we assume $\mathbf{D}$ to be implicit in the calculation of the likelihoods and pseudolikelihoods.

To prove (1) holds under the scenario (*), it is sufficient to show that the following is true for some $\epsilon > 0$

$$\frac{1}{n}(\ln \mathcal{PL}_n(\hat{\theta}_2) - \ln \mathcal{PL}_n(\hat{\theta}_1)) > \epsilon \quad (2)$$

It was shown in [3] that for any $\mathcal{M}_1$ outside of a neighbourhood N of θ_0, and $\mathcal{M}_2$ containing this neighbourhood, we can pick a $\delta > 0$ such that:

$$\frac{1}{n}(\ln \mathcal{L}_n(\hat{\theta}_2) - \ln \mathcal{L}_n(\hat{\theta}_1)) > \delta \quad (3)$$

In order to extend this result to (2), we invoke a result from [6] stating that

$$\mathcal{PL}_n(\theta) \geq d\mathcal{L}_n(\theta) + \sum_{i=1}^{d} H_i(\widetilde{P}_n) \quad (4)$$

where d is the dimensionality of the data, and $H_i(\widetilde{P}_n)$ is the Shannon entropy of the empirical distribution. It then follows that (2) holds when (3) is true.

Showing that (1) holds under the scenario (**) is equivalent to showing that the following difference is $O_p(1/n)$:

$$\frac{1}{n}|\ln \mathcal{PL}_n(\hat{\theta}_1) - \ln \mathcal{PL}_n(\hat{\theta}_2)| \quad (5)$$

Consider the difference between the full log-likelihoods:

$$\frac{1}{n}|\ln \mathcal{L}_n(\hat{\theta}_1) - \ln \mathcal{L}_n(\hat{\theta}_2)|. \quad (6)$$

We first closely follow the proof in [3] to show that the quantity in (6) is $O_p(1/n)$. Consider data drawn from a curved exponential family density $p(\mathbf{X};\theta) = h(\mathbf{X})\exp(\theta T(\mathbf{X}) - Z(\theta))$, where $\theta \in \mathbb{R}^k$ is a set of canonical parameters in the natural parameter space Θ, $T(\mathbf{X})$ is a set of sufficient statistics, and $Z(\theta)$ is a normalizing function. For a particular choice of a model $\mathcal{M}$ in this setting, the BIC can be written as $\ln\mathcal{L}_n(\mathbf{D};\hat{\theta}) - \frac{k}{2}\ln(n)$ or equivalently,

$$\sup_{\theta\in\mathcal{M}\cap\Theta} \sum_{i=1}^{n} \theta T(\mathbf{X}_i) - Z(\theta) - \frac{k}{2}\ln(n), \quad (7)$$

Note that for simplicity of notation and without loss of generality, we set $h(\mathbf{X}) = 1$. Now consider $\mathbf{T}_n = \frac{1}{n}\sum_{i=1}^{n} T(\mathbf{X}_i)$, the sample average of the sufficient statistics. We can then express (7) as

$$n \sup_{\theta\in\mathcal{M}\cap\Theta} \theta\mathbf{T}_n - Z(\theta) - \frac{k}{2}\ln(n). \quad (8)$$

Define the quantities $S_{n,i}$ and U_n as,

$$S_{n,i} \equiv \sup_{\theta_i\in\mathcal{M}_i\cap\Theta} \theta_i\mathbf{T}_n - Z(\theta_i) = \hat{\theta}_{n,i}\mathbf{T}_n - Z(\hat{\theta}_{n,i}),$$
$$U_n \equiv \theta_0\mathbf{T}_n - Z(\theta_0),$$

where $\hat{\theta}_{n,i}$ is the MLE. We now show that $S_{n,i} - U_n$ and by extension each term in (6) is $O_p(1/n)$. Since θ_0 lies in both model spaces under scenario (**),

$$S_{n,i} - U_n = (\hat{\theta}_{n,i} - \theta_0)\mathbf{T}_n - Z(\hat{\theta}_{n,i}) + Z(\theta_0) \geq 0. \quad (9)$$

Considering the Taylor expansion of Z about θ_0, we have that $Z(\hat{\theta}_{n,i}) - Z(\theta_0) = (\hat{\theta}_{n,i} - \theta_0)\nabla Z(\theta_0) + O_p(1/n)$,

where the $O_p(1/n)$ term comes from the efficiency of MLE [4]. Plugging this into (9) we get,

$$S_{n,i} - U_n = (\mathbf{T}_n - \nabla Z(\theta_0))(\hat{\theta}_{n,i} - \theta_0) + O_p(1/n). \quad (10)$$

By the Central Limit Theorem, $\mathbf{T}_n - \nabla Z(\theta_0)$ is $O_p(1/\sqrt{n})$ and by the efficiency of MLE, $\hat{\theta}_{n,i} - \theta_0$ is also $O_p(1/\sqrt{n})$. Thus, $S_{n,i} - U_n$ is $O_p(1/n)$, and we have our result.

In order to extend this result to (5), we once again invoke the result from [6] that

$$\mathcal{PL}_n(\theta) \geq d\mathcal{L}_n(\theta) + \sum_{i=1}^{d} H_i(\widetilde{P}_n) \quad (11)$$

where $H_i(\widetilde{P}_n)$ is the Shannon entropy of the empirical distribution. We see that as long $d \ll n$ (which in our setting we assume to be true), (6) being $O_p(1/n)$ implies that (5) is as well. □

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: $\sum_{V \in \mathrm{loc}(V_i, V_j; \mathcal{G}) \cap \boldsymbol{B}_{\mathrm{loc}}} \{s_V(\boldsymbol{D}; \mathcal{G}) - s_V(\boldsymbol{D}; \mathcal{G}')\}$, where $s_V(.)$ denotes the component of the score for V.*

Proof. A conditional MRF corresponding to $p(\mathbf{B} \mid \mathrm{pa}_{\mathcal{G}}(\mathbf{B}))$ for a block $\mathbf{B}$ in a CG $\mathcal{G}$ in the (conditional) exponential family has a probability distribution of the general form:

$$p(\mathbf{B} \mid \mathrm{pa}_{\mathcal{G}}(\mathbf{B}); \psi) = \quad (12)$$
$$\exp\left(\sum_{\{\mathbf{C} \in \mathcal{C}((\mathcal{G}_{\mathrm{bd}_{\mathcal{G}}(\mathbf{B})})^a) : \mathbf{C} \not\subseteq \mathrm{pa}_{\mathcal{G}}(\mathbf{B})\}} \psi_{\mathbf{C}} T(\mathbf{C}) - Z(\psi, \mathrm{pa}_{\mathcal{G}}(\mathbf{B})) \right)$$

where

$$\left\{ \psi_{\mathbf{C}} : \mathbf{C} \in \mathcal{C}((\mathcal{G}_{\mathrm{bd}_{\mathcal{G}}(\mathbf{B})})^a), \mathbf{C} \not\subseteq \mathrm{pa}_{\mathcal{G}}(\mathbf{B}) \right\}$$

is a set of canonical parameters associated with potential functions $\phi_{\mathbf{C}}$ in the CG factorization,

$$\left\{ T(\mathbf{C}) : \mathbf{C} \in \mathcal{C}((\mathcal{G}_{\mathrm{bd}_{\mathcal{G}}(\mathbf{B})})^a), \mathbf{C} \not\subseteq \mathrm{pa}_{\mathcal{G}}(\mathbf{B}) \right\}$$

is a set of sufficient statistics for $\psi_{\mathbf{C}}$, and $Z(\theta, \mathrm{pa}_{\mathcal{G}}(\mathbf{B}))$ is a normalizing function.

Assume V is in a clique $\mathbf{C}$ that contains the edge $V_i - V_j$ in $\mathcal{G}$, and let $\mathcal{G}^-$ be the edge subgraph of $\mathcal{G}$ with that edge removed. Then $p(V \mid \mathrm{bd}_{\mathcal{G}}(V))$ will only be a function of clique parameters $\psi_{\mathbf{S}}$, where $\mathbf{S} \subseteq \mathcal{C}((\mathcal{G}_{\mathrm{bd}_{\mathcal{G}}(\mathbf{B})})^a) : \mathbf{C} \not\subseteq \mathrm{pa}_{\mathcal{G}}(\mathbf{B})$ and $V \in \mathbf{S}$. All others terms in the factorization cancel by definition of conditioning. As a consequence, $p(V \mid \mathrm{bd}_{\mathcal{G}}(V))$ will be a function of $\psi_{\mathbf{C}}$.

However, after $V_i - V_j$ is removed, $\mathbf{C}$ will no longer be a clique in $\mathcal{G}^-$, by definition, but will instead decompose into two cliques, say $\mathbf{C}_1$ and $\mathbf{C}_2$. By following the above reasoning, $p(V \mid \mathrm{bd}_{\mathcal{G}^-}(V))$ will be a function of all clique parameters $\{\psi_{\mathbf{S}} : \mathbf{S} \subseteq \mathcal{C}((\mathcal{G}_{\mathrm{bd}_{\mathcal{G}}(\mathbf{B})})^a), \mathbf{C} \not\subseteq \mathrm{pa}_{\mathcal{G}}(\mathbf{B}), V \in \mathbf{S}\}$, which will include $\psi_{\mathbf{C}_1}$ and $\psi_{\mathbf{C}_2}$. Since the parameterization for $p(V \mid \mathrm{bd}_{\mathcal{G}^-}(V))$ is thus different in models for $\mathcal{G}$ and $\mathcal{G}^-$, the contribution to the score associated with this term will also be different.

Assume V is not in a clique that contains the edge $V_i - V_j$ in $\mathcal{G}$, and let $\mathcal{G}^-$ be the edge subgraph of $\mathcal{G}$ with that edge removed, as before. Then $p(V \mid \mathrm{bd}_{\mathcal{G}}(V))$ will only be a function of clique parameters $\psi_{\mathbf{S}}$, where $\mathbf{S}$ contains V, all others will cancel by definition of conditioning.

Note that since no such $\mathbf{S}$ contains the edge $V_i - V_j$ in $\mathcal{G}$, the set of cliques $\mathbf{S}$ in $\mathcal{G}$ is the same as the set of cliques $\mathbf{S}$ in $\mathcal{G}^-$. Moreover, since $\mathcal{G}^-$ is an edge subgraph of $\mathcal{G}$, no new cliques are introduced. As a result, $p(V \mid \mathrm{bd}_{\mathcal{G}^-}(V))$ will be parameterized by the same set of $\psi_{\mathbf{S}}$ in the model for $\mathcal{G}^-$ as it was in the model for $\mathcal{G}$.

Our conclusion then follows because, by properties of the exponential family, the sufficient statistics for a clique parameter $\psi_{\mathbf{S}}$ are functions of only $\mathbf{S}$. Since draws from $p(\mathbf{S})$ are fixed, the estimates for $\psi_{\mathbf{S}}$ will coincide if the data is evaluated under the model for $\mathcal{G}$, and the model for $\mathcal{G}^-$. Furthermore, the number of parameters in $p(V \mid \mathrm{bd}_{\mathcal{G}}(V))$ and $p(V \mid \mathrm{bd}_{\mathcal{G}^-}(V))$ is the same. This implies the score contribution for $p(V \mid \mathrm{bd}_{\mathcal{G}}(V))$ in $\mathcal{G}$ will equal the score contribution of $p(V \mid \mathrm{bd}_{\mathcal{G}^-}(V))$ in $\mathcal{G}^-$. The only terms remaining in the score difference between $\mathcal{G}$ and $\mathcal{G}'$ are then local scores for $V \in \mathrm{loc}(V_i, V_j; \mathcal{G})$.

This implies the conclusion. □

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.*

Proof. Under the restrictions listed above, the only changes allowed are edge deletions or additions of the form $L_i - L_j$, $A_i - A_j$, $Y_i - Y_j$, $L_i \to A_j$, $L_i \to Y_j$, $A_i \to Y_j$.

Consider an edge deletion $V_i - V_j$ in $\mathcal{G}$, giving rise to a graph $\mathcal{G}'$. Notice that boundaries of V_i and V_j have changed. Thus by the local Markov property on chain graphs, $\mathcal{G}$ and $\mathcal{G}'$ must imply different conditional inde-

pendences. Concretely, $\mathcal{G}$ implies:

$$V_i \perp\!\!\!\perp \mathbf{V} \setminus \mathrm{cl}_{\mathcal{G}}(V_i) \mid \mathrm{bd}_{\mathcal{G}}(V_i)$$
$$V_j \perp\!\!\!\perp \mathbf{V} \setminus \mathrm{cl}_{\mathcal{G}}(V_j) \mid \mathrm{bd}_{\mathcal{G}}(V_j)$$

while $\mathcal{G}'$ implies:

$$V_i \perp\!\!\!\perp \mathbf{V} \setminus (\mathrm{cl}_{\mathcal{G}}(V_i) \setminus V_j) \mid \mathrm{bd}_{\mathcal{G}}(V_i) \setminus V_j$$
$$V_j \perp\!\!\!\perp \mathbf{V} \setminus (\mathrm{cl}_{\mathcal{G}}(V_j) \setminus V_i) \mid \mathrm{bd}_{\mathcal{G}}(V_j) \setminus V_i$$

We can similarly show that an edge deletion $V_i \to V_j$ also implies different conditional independences in $\mathcal{G}$ and $\mathcal{G}'$. Thus, in general, an edge deletion or addition in our search space gives rise to graphs that are not Markov equivalent and hence, reside in 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.*

Proof. The algorithm begins with a complete conditional MRF that contains the true underlying distribution. We are guaranteed that the truth is contained in every state through the entirety of the algorithm by the following argument. Consider the first edge deletion performed by GNS to a conditional MRF that does not contain the true model. It follows from consistency of the PBIC that any such deletion would decrease the score. Choosing such an edge deletion would contradict the greediness of the algorithm.

Now assume the algorithm stops at a sub optimal conditional MRF $\mathcal{G}$ that contains the truth but has more parameters than the true model $\mathcal{G}^*$. We know there exists a series of single edge deletions in $\mathcal{E}_{\mathcal{N}}$ that takes us from $\mathcal{G}$ to $\mathcal{G}^*$. By Lemma 3, each of these edge deletions yield graphs in separate equivalence classes. It follows then from the consistency of the PBIC that each of these edge deletions strictly increases the score (each edge deletion yields a smaller model containing the truth) and thus, a local optimum found by greedily maximizing the PBIC corresponds to finding the global optimum $\mathcal{G}^*$. □

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

Proof. By consistency of GNS, each conditional MRF returned for **L**, **A**, and **Y** corresponds to the true model. The union of these will then produce the true CG on **V**. □

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

Proof. Each of the homogenous procedures described above can be decomposed into a series of single edge deletions that we have shown to be consistent. □

References

[1] 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.

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

[3] Dominique M. A. Haughton. On the choice of a model to fit data from an exponential family. *Annals of Statistics*, 16(1):342–355, 1988.

[4] Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 221–233. University of California Press, 1967.

[5] 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.

[6] Alexander Mozeika, Onur Dikmen, and Joonas Piili. Consistent inference of a general model using the pseudolikelihood method. *Phys. Rev. E*, 90:010101, 2014.

[7] Elizabeth L. Ogburn, Ilya Shpitser, and Youjin Lee. Causal inference, social networks, and chain graphs. *arXiv preprint arXiv:1812.04990*, 2018.

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

[9] Daniel Westreich, Stephen R. Cole, Jessica G. Young, Frank Palella, Phyllis C. Tien, Lawrence Kingsley, Stephen J. Gange, and Miguel A. Hernán. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident aids or death. *Statistics in Medicine*, 31(18):2000–2009, 2012.