JMLR Workshop and Conference Proceedings 6:165–176 NIPS 2008 workshop on causality

# Structure Learning in Causal Cyclic Networks

**Sleiman Itani*** SSOLOMON@MIT.EDU
*77 Massachusetts Ave,*
*32-D760*
*Cambridge, MA, 02139, USA*

**Mesrob Ohannessian*** MESROB@MIT.EDU
*77 Massachusetts Ave,*
*32-D740*
*Cambridge, MA, 02139, USA*

**Karen Sachs*** KAREN.SACHS@STANFORD.EDU
*Stanford University School of Medicine,*
*269 Campus Drive*
*Stanford, CA, 94305, USA*

**Garry P. Nolan,** GNOLAN@STANFORD.EDU
*Stanford University School of Medicine,*
*269 Campus Drive*
*Stanford, CA, 94305, USA*

**Munther A. Dahleh** DAHLEH@MIT.EDU
*77 Massachusetts Ave,*
*32-D734*
*Cambridge, MA, 02139, USA*
* *These authors contributed equally to this work.*

**Editor:** Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

## Abstract

Cyclic graphical models are unnecessary for accurate representation of joint probability distributions, but are often indispensable when a causal representation of variable relationships is desired. For variables with a cyclic causal dependence structure, DAGs are guaranteed not to recover the correct causal structure, and therefore may yield false predictions about the outcomes of perturbations (and even inference.) In this paper, we introduce an approach to generalize Bayesian Network structure learning to structures with cyclic dependence. We introduce a structure learning algorithm, prove its performance given reasonable assumptions, and use simulated data to compare its results to the results of standard Bayesian network structure learning. We then propose a modified, heuristic algorithm with more modest data requirements, and test its performance on a real-life dataset from molecular biology, containing causal, cyclic dependencies.

## 1. Introduction

Bayesian network models encode probabilistic relationships among random variables, providing a framework for tasks such as inference and decision making. In some settings, it is useful for model edges to represent probabilistic dependence resulting from causal mechanisms. This is the case when the goal is structure recovery for the sake of revealing causal interactions for prediction of perturbation effects in some domain, for instance, when learning the structure of molecular pathways from biological measurements.

Causal Bayesian network models have been described (Pearl, 2000), relying on the *framework of causation*, which enables causal interpretation under proper assumptions (Spirtes et al., 1993). These models may be learned from *observational data*, i.e. passive observations of the domain. However, such methods yield entire equivalence classes, leaving the causal direction of many edges unknown. A solution to this problem is offered by the *framework of intervention*, where interventions effectively override variables, and halt the influence of the network on them, enabling the use of *interventional* or *experimental* data (Pearl, 1995) and (Pearl, 2000). In this framework, it is possible to ask: "how can the graphical structure of the causal model be recovered from observational and experimental data?"

Research in Bayesian networks has predominantly focused on directed acyclic graphs (DAGs), even when the acyclicity assumption is knowingly violated (Friedman et al., 2000). Within that context, solutions to this question abound, e.g. (Cooper and Yoo, 1999). In cyclic domains, DAGs represent an inaccurate causal structure, consequently, prediction of perturbation effects will fail, as in Figure 1. To avoid these inaccuracies, a representation which encompasses cycles must be employed.

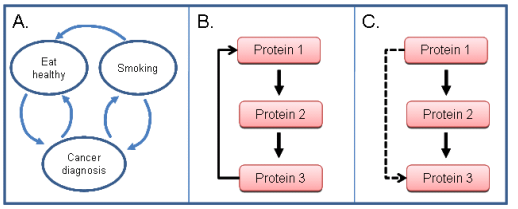


Figure 1: **Cyclic causal networks**. **A.** Risk assessment network for predicting the effect of behavior interventions. Smoking positively influences diagnosis of lung cancer, while eating healthy does so negatively. A cancer diagnosis may influence eating and smoking choices, though cessation of smoking can deteriorate eating habits. **B.** In protein networks, feedback loops are ubiquitous modes of positive and negative regulation of biological processes. **C.** A DAG representation of the cyclic structure in B. The dotted line indicates an incorrectly oriented edge: perturbing protein 3 would inaccurately be assessed as having no effect on proteins 1 and 2.

BN models with directed cyclic graphs (DCGs), though inherently possible, (Pearl, 1988), had unclear interpretation and applicability (Spirtes et al., 1993). Cycles also occur in an alternative modeling paradigm called structural equation models (SEMs), which model functional dependence directly. Key developments (Spirtes, 1995) and (Koster, 1996), endowed DCGs with some of the properties of their DAG counterparts, and it was also shown that some SEMs are amenable to the same analysis (Spirtes, 1995; Pearl and Dechter, 1996). Based on these, (Richardson, 1996) established an algorithm for discovering a partial structure on DCGs. Recently, (Lacerda et al., 2008) provided an alternative algorithm. Both procedures lie in the framework of causation, and use solely observational data to output equivalence classes, rather than a single DCG.

In this paper, we are interested in modeling cycles, yet tapping into the power of experimental data. At the extreme of exhaustive interventions, the problem appears trivial. However,

discovering structure by such brute force is a daunting task, and in truth one is constrained by the number and type of interventions at hand. We address the problem through the following contributions:

- In Section 2, we give a novel formalization of cyclic networks by characterizing them locally with stochastic kernels, which bridge the SEM context with that of BNs by replacing deterministic equations with exogenous variables by a direct probabilistic description. We call the resulting models generalized Bayesian networks (GBNs). The framework of intervention extends directly to such a description, resulting in causal GBNs (CGBNs).

- In Section 3, we prove that interventions allow us to discover descendants and children. Such discovery is robust, in that in general it does not result in false discovery and, given natural properties, it always succeeds as the size of the data grows to infinity. Interventions can affect either the *abundance* or the *activity* of variables (corresponding to ingoing or outgoing edges, respectively). However, in this work, we assume the *activity* of a perturbed variable is affected. We elaborate in Section 2.3.

- In Section 4, we cast these results into an algorithm for structure learning. Rather than searching over all causal interactions by brute force, we first discover cycle breakers. Upon intervention on these quantities, we reduce the task into an acyclic problem which can be learned generically. Finally we close cycles to recover the cyclic structure. We illustrate these results on synthetic data with 14 nodes, 2 cycles and 3 interventions.

- In Section 5, we develop a modified heuristic algorithm for the structure learning from more limited data, containing only one perturbation per sample. This algorithm is inspired by our previous one, and is motivated by limitations on experiment technologies. We illustrate the usefulness of this algorithm by studying a biological dataset of 11 variables from the MAPK/AKT pathway (CYTO) (Sachs et al., 2005).

Finally, a related research area is that concerned with structure learning with time course data. In this case, alternative representations exist in the form of dynamic Bayesian networks (DBNs) (Friedman et al., 1999) and continuous-time Bayesian networks (CTBNs) (Nodelman et al., 2002, 2003). These models represent cycles by 'unrolling' them in time. As with other efforts to learn static representations of underlying dynamic systems (Friedman et al., 2000; Sachs et al., 2005), what we propose here can be interpreted as learning a DBN or CTBN in the absence of time-course data, or from single time-point data (constituting a snapshot of a dynamic system).

## 2. Problem formulation

### 2.1 Generalized Bayesian networks

**Definition 1 (Generalized Bayesian network)** *We define a* generalized Bayesian network *(GBN) as a pair $(G,F)$, where $G$ is a directed graph $G=(V,E)$ and $F$ is a set of stochastic kernels (conditional probability tables) $f_i : \mathcal{X} \times \mathcal{X}^{|\pi_i|} \longrightarrow \mathbf{R}_+$ indexed by all nodes $i \in V$, for a finite set $\mathcal{X}$. Here, $\pi_i$ is the set of parents of $i$ in $G$. With each node $i$ of the GBN we associate a random variable $X_i$. In this paper, we restrict ourselves to discrete random variables taking values in a common alphabet $\mathcal{X}$.* [1] *The GBN then* induces *a joint distribution on $X_1, \cdots, X_N$ satisfying the following* characterizations*:*

1. Although we restrict ourselves to discrete variables, this is in general not restrictive since any continuous variable can approximated arbitrarily well by a discrete variable.

*(i). Local characterization:*

$$\mathbb{P}(X_i = x_i, X_{\pi_i} = x_{\pi_i}) = \mathbb{P}(X_{\pi_i} = x_{\pi_i}) f_i(x_i; x_{\pi_i}), \qquad \forall i \in V. \tag{1}$$

*(ii). Independence under d-separation: Given any two nodes i and j in G, if i and j are d-separated (Pearl, 1988) by a set $Z \subset V$, then $X_i$ and $X_j$ given $\{X_k, k \in Z\}$.*

We make the following assumption:

Assumption *[Existence and Uniqueness] For every F that we consider, there exists a unique induced (global) joint distribution that satisfies all the local characterizations in Equation (1).*

Since *GBN*'s are generalizations of BN's to the cyclic case, the previous assumption doesn't hold for any graph $G$ and stochastic kernels $\{f_i\}$. This is just like the fact that a dynamic system with feedback (cycle) is not necessarily causal even if all of the subsystems are causal. Of course, it is expected that in the applications of interest, the variables measured *do* come from a unique underlying joint distribution.. Another view of this assumption is that it is the same as the one in the case of the Gibb's sampler: for the sampling to guarantee convergence, a unique joint that is compatible with the given conditionals must exist.

When the graph of a GBN is acyclic, the product of all the stochastic kernels gives a valid joint distribution satisfying (1). Thus, by uniqueness, an acyclic GBN reduces to a BN:

$$\mathbb{P}(X_1 = x_1, \cdots, X_N = x_N) = \prod_{i \in V} f_i(x_i; x_{\pi_i}). \tag{2}$$

### 2.2 Causal generalized Bayesian networks

Let an *intervention* $(I, \xi)$ be a pair, where $I \subset V$ is a subset of the nodes of a graph $G$, and $\xi \in \mathcal{X}^{|I|}$ is a tuple of values in an alphabet $\mathcal{X}$.

**Definition 2 (Causal generalized Bayesian network)** *We define a* causal generalized Bayesian network *(CGBN) as a GBN with which we associate a collection of joint distributions $\mathbb{P}_{(I,\xi)}$ indexed by all interventions $(I, \xi)$, for each of which it satisfies:*

$$\mathbb{P}_{(I,\xi)}(X_i = x_i, X_{\pi_i \setminus I} = x_{\pi_i \setminus I}) = \mathbb{P}_{(I,\xi)}(X_{\pi_i \setminus I} = x_{\pi_i \setminus I}) f_i(x_i; x_{\pi_i \setminus I}, \xi_{\pi_i \cap I}), \qquad \forall i \in V. \tag{3}$$

When $\xi$ is implicit we only use $I$ as subscript, and when $I = \varnothing$ we drop the subscript altogether. Below, we provide more intuition about this definition. Meanwhile, we extend the assumption of existence and uniqueness to CGBNs by taking it to hold for every intervention $(I, \xi)$. With this, an acyclic CGBN reduces to a causal BN, in the sense of interventions (Pearl, 2000):

$$\mathbb{P}_{(I,\xi)}(X_1 = x_1, \cdots, X_N = x_N) = \prod_{i \in V} f_i(x_i; x_{\pi_i \setminus I}, \xi_{\pi_i \cap I}). \tag{4}$$

### 2.3 $\sigma$-$\mu$ characterization

By carefully examining Equation (3), we can see how interventions effectively decouple nodes into *seen* and *measured* values. Just as in the do-calculus of Pearl, the intervention value supersedes the node variable itself as far as its influence on the network goes, and can thus be interpreted as what is (internally) *seen* by all descendants. The value of the *seen* variable is determined solely by the intervention. However, and this is in contrast to traditional intervention models, we (externally) *measure* or *observe* the value (i.e. abundance) of the intervention variables. These can be thought of as shadow copies, which are still influenced by the network

*but no longer influence it*, because its activity is externally set by the intervention. This formulation is motivated by some inhibition models in molecular biology, where the inhibitors do not change the amount of a given protein but rather halt its activity. Thus the correct modeling of this situation is to separate the inhibited node from its children. The $\sigma$-$\mu$ characterization simply does that while staying in the framework of probability theory. All of our results extend to the case when measured values are lost, by eliminating the variables intervened at.

We can capture this decoupling via an explicit characterization which reduces a CGBN with an intervention to a GBN. In particular, given a CGBN $(G,F)$ describing $N$ variables and an intervention $(I,\xi)$, one can construct a GBN $(G',F')$ which describes $N+|I|$ variables, such that the restriction to the first $N$ of the variables has a joint distribution evaluating to $\mathbb{P}_{(I,\xi)}$. We call this construction the $\sigma$-$\mu$ characterization of a CGBN. We do not elaborate on this further, and leave its illustration to the second example below.

### 2.4 Examples

#### 2.4.1 CYCLE WITH 2 NODES

Consider the GBN with binary-valued variables $X_1$ and $X_2$ described in Figure 2. The local characterizations of the joint distribution $\mathbb{P}$ induced by the GBN are as follows: $\mathbb{P}(X_1 = x_1, X_2 = x_2) = \mathbb{P}(X_2 = x_2) f_1(x_1; x_2)$, and $\mathbb{P}(X_1 = x_1, X_2 = x_2) = \mathbb{P}(X_1 = x_1) f_2(x_2; x_1)$, for all binary configurations of $x_1$ and $x_2$. Under the proper choice of $f_1$ and $f_2$, these yield linearly independent equations, in which case a distribution satisfying the local characterizations exists and is unique.

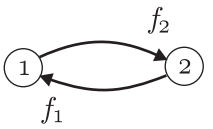


Figure 2: A GBN with two nodes.

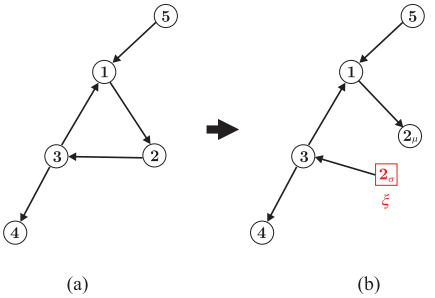


Figure 3: A CGBN with an intervention at node 2.

#### 2.4.2 BREAKING CYCLES

Consider the CGBN described in Figure 3a. In Figure 3b, we illustrate what happens when node 2 is intervened at. We use the $\sigma$-$\mu$ characterization, and represent the seen node with a $\sigma$ subscript and the measured node with a a $\mu$ subscript. Note how node $2_\mu$ is effectively a leaf under intervention. As such, the resulting graph is a DAG. It follows that, the product of all the $f_i$'s is a valid characterization, and by uniqueness it is the distribution induced by the CGBN under the intervention. The resulting network is thus exactly equivalent to a BN. We say that the cycle has been *broken*. This notion, in more generality, will be used throughout our algorithm (Section 4).

## 3. Interventions and Descendent Detection

We now introduce analytical results which we subsequently use to justify the correctness of our algorithm for structure learning. For conciseness, we state and prove only the forward

direction of the results. The converses hold under some natural properties of the network and interventions. Please see supporting materials for additional proofs and associated assumptions.

**Theorem 3** *Consider a CGBN, and let the existence and uniqueness assumption hold. Intervene at a single node i, that is let* $(I,\xi) = (i,\xi_i)$ *and consider a node* $j$*. If* $j$ *is not a descendant of* $i$ *then* $\mathbb{P}(X_j = x_j) = \mathbb{P}_{(i,\xi i)}(X_j = x_j)$ *for all* $x_j \in \mathcal{X}$*.*

**Proof** Partition $V$ into two: $V_i$ (nodes that are descendants of $i$, including $i$) and $\overline{V_i}$ (nodes that are not descendants of $i$). Consider the network restricted to $\overline{V_i}$, by restricting the graph. Since there are no incoming edges from $V_i$ to $\overline{V_i}$, we can also restrict $F$ to contain only $f_j$, $j \in \overline{V_i}$. Since none of the local characterizations of the distribution induced by the restricted network depend on the intervention, and by the uniqueness of the solution, the restricted distribution is unchanged. Thus the marginal distributions of all $j \in \overline{V_i}$ is unchanged. ■

In other words, Theorem 3 states that if a node $j$ experiences a change in marginal distribution when $i$ is intervened at, then it is a descendant of $i$. As mentioned, the converse also holds under proper assumptions, detailed in the supporting materials. One of these assumptions states that a child variable must be sensitive to perturbations imposed upon its parent variables, an assumption which may in general be violated, particularly if the network compensates in the face of perturbations. Such *insensitive* descendants may still be detectable with the use of multiple perturbations.

**Theorem 4** *Consider a CGBN, an intervention* $(I^1,\xi^1)$*, and an incremental intervention* $(I^2,\xi^2)$ *by a single node i, as in* $I^2 \setminus I^1 = \{i\}$*. Let the existence and uniqueness assumption hold. Define* $\mathbb{P} := \mathbb{P}_{(I^1,\xi^1)}$ *and* $\mathbb{Q} := \mathbb{P}_{(I^2,\xi^2)}$*. Consider a node* $j$ *and let* $\tilde{\pi}_j = \pi_j \setminus I^2$*. If* $j$ *is not a child of* $i$ *then* $\mathbb{P}(X_j = x_j | X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j}) = \mathbb{Q}(X_j = x_j | X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j})$ *for all* $x_j \in \mathcal{X}$ *and* $x_{\tilde{\pi}_j} \in \mathcal{X}^{|\tilde{\pi}_j|}$*.*

**Proof** We shall split the parents of $j$ into three groups: $i$ itself if it is a parent, the never-intervened-at parents $\tilde{\pi}_j$, and the always-intervened-at parents $\hat{\pi}_j$. When $j$ is not a child of $i$ the inclusion pattern for the parents of $j$ in the local characterization is unchanged. Hence:

$$\begin{aligned}
\mathbb{P}(X_j = x_j | X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j}) &= \frac{\mathbb{P}(X_j = x_j, X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j})}{\mathbb{P}(X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j})} = f_j(x_j; x_{\tilde{\pi}_j}, \xi^1_{\hat{\pi}_j}), \\
\mathbb{Q}(X_j = x_j | X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j}) &= \frac{\mathbb{Q}(X_j = x_j, X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j})}{\mathbb{Q}(X_{\tilde{\pi}_j} = x_{\tilde{\pi}_j})} = f_j(x_j; x_{\tilde{\pi}_j}, \xi^2_{\hat{\pi}_j}).
\end{aligned}$$

But since $\xi^1$ and $\xi^2$ agree on $\hat{\pi}_j$, the claim follows. ■

In other words, Theorem 4 states that if a node $j$ experiences a change in marginal conditional distribution given the never-intervened-at parents $\tilde{\pi}_j$ when $i$ is intervened at, then it is a child of $i$. Again, the converse also holds under proper assumptions.

## 4. Algorithm for structure learning

Consider a CGBN from which we can sample both observational and experimental data, from an intervention set $I$ and its subsets. Assume that $I$ is 'rich', in the sense that it has at least one representative node from every cycle in the underlying graph. The following algorithm effectively guides the experimental procedure (or uses previously collected data) and recovers

the CGBN's structure. In what follows, we elaborate the subroutines that are used, and show correctness.

**Algorithm: Learn CGBN structure**

0: Start with a CGBN and an intervention set $I$.
1: [Probing experiments] Collect sets of i.i.d. samples under no-intervention and single-intervention data, i.e. when node $i$ is intervened at, for each $i$ in $I$.
2: Call subroutine 'detect descendants' to recover descendant information for all nodes in $I$.
3: Identify the minimal subset of nodes in $I$ which are sufficient to break all cycles, and denote it by $I_{\mathcal{C}}$.
4: [Cycle-breaking experiment] Collect i.i.d. samples when all nodes in $I_{\mathcal{C}}$ are intervened at.
5: Recover an embedded DAG.
6: [Leave-one-out experiments] Collect sets of i.i.d. samples when nodes in $I_{\mathcal{C}} \setminus \{i\}$ are intervened at, for each $i \in I_{\mathcal{C}}$.
7: Call subroutine 'detect children' to recover child information for all nodes in $I_{\mathcal{C}}$.
8: Recover all missing edges in the DAG, and complete the DCG structure of the CGBN.

The following is the subroutine that obtains descendant information based on no-intervention and single-intervention i.i.d. data. The correctness of the subroutine follows from Theorem 3 and the convergence of empirical distributions, since non-descendants will exhibit no change of marginal, whereas descendants will. The choice of distance is not critical, and thresholding can be automated.

**Subroutine: Detect descendants**

0: Start with sets of $n$ i.i.d. samples generated by a CGBN, under no interventions as well as single-interventions at each $i$ in $I$. Initialize a binary $|V| \times |I|$ descendant information matrix.
1: For each $j \in V$:
2: Compute $\hat{\mathbb{P}}^n(X_j)$, the empirical marginal of $X_j$ under no interventions.
3: For each $i \in I$:
4: Compute $\hat{\mathbb{P}}^n_i(X_j)$, the empirical marginal of $X_j$ under the single-intervention $i$.
5: Evaluate some distance between $\hat{\mathbb{P}}^n(X_j)$ and $\hat{\mathbb{P}}^n_i(X_j)$.
6: If the distance exceeds a threshold, mark $j$ as a descendant of $i$.
7: Next $i$.
8: Next $j$.
9: Compute the transitive closure of the descendant information matrix, and return it.

$I_{\mathcal{C}}$ can then be identified as the set of all self-descendants. Since the intervention set $I$ has at least one node from each cycle in the underlying graph, $I_{\mathcal{C}}$ constitutes a cycle-breaking intervention set, meaning that if all nodes in $I_{\mathcal{C}}$ are intervened at, the CGBN behaves like a BN. Thus with i.i.d. data obtained as such, we can recover the corresponding embedded DAG using generic BN structure learning, which we do not elaborate further on. Note that $I$ itself is a cycle-breaking intervention set, the merit here being that $I_{\mathcal{C}}$ can be much smaller.

Note that the only edges that are in the underlying graph but are missing from the embedded DAG are those from cycle breakers to their children. The following subroutine obtains a child information matrix, based on $I_{\mathcal{C}}$-intervention and leave-one-out from $I_{\mathcal{C}}$ intervention i.i.d. data. Once this information is obtained, all cycles can be closed in a straightforward fashion, recovering the underlying structure. Once again, the correctness of the subroutine follows from Theorem 4 and the convergence of empirical distributions, since only children will exhibit a change in marginal conditional.

**Subroutine: Detect children**

0: Start with the recovered DAG, and sets of $n$ i.i.d. samples generated by the CGBN, under $I_{\mathcal{C}}$-intervention as well as leave-one-out interventions, i.e. on $I_{\mathcal{C}} \setminus \{i\}$ for each $i$ in $I_{\mathcal{C}}$. Initialize a binary $|V| \times |I_{\mathcal{C}}|$ child information matrix. Denote by $\hat{\pi}_j$ the parents of node $j$ according to the recovered DAG.
1: For each $j \in V$:
2: For each $\alpha \in \mathcal{X}^{|\hat{\pi}_j|}$:
3: Compute the empirical marginal conditional $\hat{\mathbb{P}}^n_{I_{\mathcal{C}}}(X_j | X_{\hat{\pi}_j} = \alpha)$, call it $\mathbb{Q}_1$.
4: For each $i \in I_{\mathcal{C}}$:
5: Compute the empirical marginal conditional $\hat{\mathbb{P}}^n_{I_{\mathcal{C}} \setminus \{i\}}(X_j | X_{\hat{\pi}_j} = \alpha)$, call it $\mathbb{Q}_2$.
6: Evaluate some distance between $\mathbb{Q}_1$ and $\mathbb{Q}_2$.
7: If the distance exceeds a threshold, mark $j$ as a child of $i$.
8: Next $i$.
9: Next $\alpha$.
10: Next $j$.
11: Return the completed child information matrix.

To illustrate the algorithm, we simulated a GBN that has fourteen variables,shown in Figure 4, each with three states $\mathcal{X} = \{0,1,2\}$, two cycles $5 \to 6 \to 7 \to 5$ and $8 \to 9 \to 10 \to 11 \to 8$, and nodes 7, 8 and 10 available for intervention. The stochastic kernels were sampled continuously from the 3-simplex. The simulation was performed using Gibbs-like sampling (Chou et al., 1991; Sharma et al., 1989), and up to 4000 data points were sampled for every required intervention.

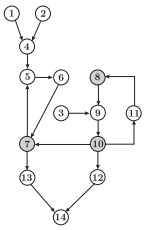

Figure 4: Test network, recovered exactly by GBN learning algorithm

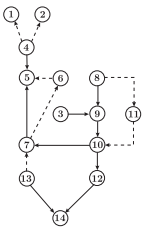

Figure 5: Best network recovered by BN structure learning

GBN algorithm

| Data | Correct | Inverted | Added |
|---|---|---|---|
| 1000 | 14 | 0 | 0 |
| 2000 | 15 | 0 | 0 |
| 4000 | 16 | 0 | 0 |

BN structure learning

| Data | Correct | Inverted | Added |
|---|---|---|---|
| 1000 | 9 | 3 | 0 |
| 2000 | 9 | 7 | 0 |
| 4000 | 12 | 4 | 2 |

Figure 6: Performance tables

In the tables of Figure 6, we compare the performance of our algorithm to a plain BN structure learning algorithm for the various data sizes. In particular, the tables document the number of true edges that the algorithms uncover, the number of reversed edges that they give, and the number of edges that they add but which are absent in the original graph. Observe that the GBN algorithm recovers the network exactly with 4000 data points. The comparison is inherently unfair, because BN structure learning does not handle cycles, but the emphasis here is on illustrating the type of pitfalls in using BNs to capture data that is generated by a GBN. Using the best recovered DAG in Figure 5, for instance, will mistakenly predict that an intervention at node 9 will not affect node 8.

## 5. Single Perturbations

In this section we introduce an algorithm that is inspired by our previous one but doesn't require data with multiple simultaneous perturbations. Due to practical considerations, sometimes multiple inhibition data-sets are not available. This is why we are interested in an algorithm that can

recover the causal structure (even when it's cyclic) without the need for multiple simultaneous perturbations. We assume that the interventions that are available are activity interventions, and so the amount of the variable $x$ can be measured when $x$ is intervened at. The algorithm we have when such perturbations are available is as follows:

**Algorithm: Learn CGBN structure without multiple simultaneous perturbations**

0: Start with a CGBN and an intervention set $I$.

1: [Probing experiments] Collect sets of i.i.d. samples under no-intervention and single-intervention data, i.e. when node $i$ is intervened at, for each $i$ in $I$.

2: Call subroutine 'detect descendants' to recover descendant information for all nodes in $I$.

3: Identify the subset of all nodes in $I$ which are in cycles, and denote it by $I_C$.

4: Use a regular CBN learning algorithm to recover an approximation of the structure of the causal relations. This is done with the standard structure learning algorithm using the complete dataset, as in Sachs et al. (2005).

5: For every variable $i$ in $I_C$:

   a- Recover the paths from $i$'s descendants in the cycle back to it using BN learning on the data where $i$ was perturbed. As in the original algorithm, this recovers the linearized structure with the perturbed node as a leaf.

   b- Overwrite the paths from $i$'s descendants in the BN approximate graph. This step may alter the parent set of $i$ as well as the direction of edges among $i$'s ancestors. Because the approximate graph is expected to have incorrect edge directionality imposed by the cycles, the graph under perturbations is considered more accurate.

6: Call subroutine 'detect children' to recover child information for all nodes in $I_C$. Use the data with no perturbations and the data with $i$ inhibited for all $i \in I_C$.

7: Recover all missing edges in the DAG, and complete the DCG structure of the CGBN. This proceeds as in the original algorithm, using only the observational data to detect direct edges and indirect paths from each variable in $I_C$ to its descendants.

This algorithm is a heuristic, although it inherits some of the intuition and reasoning of our previous algorithm: It recovers the structure of every cycle by first breaking it and finding its partial structure. To illustrate the performance of this algorithm, we applied it to a real data set from the MAPK/AKT pathway (Sachs et al., 2005).

## 6. Results from the CYTO dataset

The heuristic algorithm from Section 5 was applied to the CYTO dataset (Sachs et al., 2005), a real-life dataset of eleven protein measurements, which employs single perturbations (per sample), including three activity inhibitors and one abundance inhibitor. Model results (figure 6) show the edges from regular BN structure learning in blue (solid lines), novel edges resulting from the GBN approach in purple (broken lines). To assess this model's accuracy in representing the true underlying causal structure, as compared to the original model, we turned to the biological literature. There are seven edges unique to the GBN model, of which three represent canonical, well established causal connections that were completely missed by standard BN structure learning efforts. One of these, the connection between PIP2 and Akt, our model represents somewhat inaccurately, shifting the canonical edge (PIP3 $\rightarrow$ Akt). PIP2 and PIP3 are precursors of each other, so this edge incorrectly assigns the parent of Akt as the precursor of the actual parent, perhaps due to confounding effects of the dynamics of the system (i.e. PIP2 abundance may more accurately represent the quantity of PIP3 that influenced the current level of Akt, see Itani et al. (2009). An additional perturbation, or a more idealized one, may have helped resolved this inaccuracy. It can be argued that the GBN model with the shifted edge comes closer to representing the true structure than the BN model that fails to represent this interaction all together. Another canonical edge present only in the GBN model is PKC $\rightarrow$

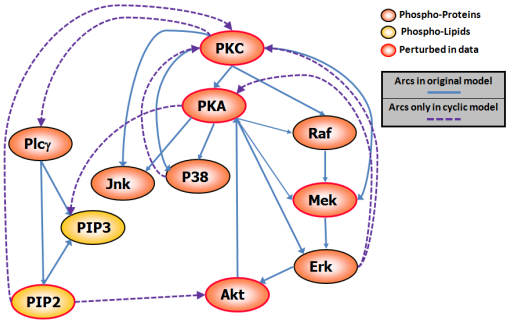


Figure 7: **Application of the heuristic algorithm to real-life protein dataset.** Application of the heuristic structure learning algorithm to this dataset from Sachs et al. (2005) yields this cyclic structure. Edges found in the original graph, resulting from standard Bayesian network structure learning, are in blue (solid lines), edges unique to the GBN result are in purple (broken lines). Several of the cycles in this result structure are supported by literature findings (see text).

Plc$\gamma$, known in the classic literature, but in the reverse orientation. While this may be an inaccuracy in direction of the edge, the data clearly support this connection (with the Plc$\gamma$ distribution strongly affected by PKC perturbation), and it has been reported by previous studies (Xu et al., 2001; Quinlan et al., 2003), leading us to believe it is a correct edge. Like its BN counterpart, the GBN model misses the edge in the Plc$\gamma \rightarrow$ PKC direction, but unlike the BN model, it successfully represents the dependence between these two proteins. Finally, the canonical edge (PIP2 $\rightarrow$ PKC) is missed by the BN model but correctly represented in the GBN model. For these canonical edges, the GBN model is somewhat imperfect but nevertheless strongly outperforms the BN model.

Of the remaining four edges, both p38 $\rightarrow$ PKC and PKA $\rightarrow$ PIP3 are supported by previous literature findings (Shimizu et al., 1999; Deming et al., 2008). We did not find specific evidence for the edges from Erk to PKC and PKA, though several studies report feedback on PKA and PKC, with potential roles for Erk (Geritsa et al., 2008). Although confirmation of all model results requires experimental validation, comparison to literature studies indicates a clear improvement in accuracy for the GBN model. Additionally, the GBN model improves on the BN result by *accurately representing all causal connections and conditional independencies found in the data*, something the standard BN model is unable to achieve.

## 7. Conclusion and future work

In this paper we reviewed previous work in incorporating both causality and cyclic structure within the context of Bayesian networks. We then presented the formalism of generalized BNs, which preserves only the local characterizations with stochastic kernels, applying it equally well to the cyclic case, under an existence and uniqueness assumption for the joint distribution. In the acyclic case, this reduces to BNs. The framework of interventions easily extends to this formalism, resulting in causal GBNs. We present an algorithm that uses no-intervention and single-intervention data to detect cycle breakers, then uses multiple simultaneous interventions to learn an embedded DAG, close cycles, and recover the underlying DCG. This algorithm relies on a minimal set of perturbations. We illustrate the procedure via a numerical example. Finally, we present a modified algorithm with more modest, one-intervention-at-a-time data requirements

and demonstrate its performance on a real-life biological dataset, successfully recovering many known connections, and strongly outperforming standard structure learning with respect to recovery of the known causal structure. This work can be extended in several directions. We are currently expanding its application to biological data by extending the algorithm to one which explicitly handles the imperfect specificity and efficacy of biological inhibitors. A more theoretical direction is that of relating snapshot structure embodied in GBNs to that of underlying time-dynamics. For that, one needs to start with a dynamic hypothesis of data generation, e.g. CTBNs, stochastic differential equations, etc. Conditions under which the static and dynamic structures coincide would further motivate the current paradigm.

## References

C. Chou, Bentler, P.M. Satorra, A. Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology*, 44, 347-357.(1991)

G. Cooper and C. Yoo (1999). Causal discovery from a mixture of experimental and observational data. *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 116–125.

P. B. Deming, S. L. Campbell, L. C. Baldor, and A. K. Howe (2008). Protein Kinase A Regulates 3-Phosphatidylinositide Dynamics during Platelet-derived Growth Factor-induced Membrane Ruffling and Chemotaxis. 10.1074 *J. Biol Chem*

N. Friedman, K. Murphy, and S. Russell (1999). Learning the structure of dynamic probabilistic networks. *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 139–147.

N. Friedman, N. Linial, I. Nachman, and D. Pe'er (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, vol. 7, pp 601–620.

N. Geritsa, S. Kostenkoa, A. Shiryaeva, M. Johannessena and U. Moens (2008). Relations between the mitogen-activated protein kinase and the cAMP-dependent protein kinase pathways: Comradeship and hostility. 20(9):1592-607. *Cellular Signaling*

S. Itani, K. Sachs, J. Fitzgerald, L. Wille, B. Schoeberl, G. Nolan and M. Dahleh (2009). Single timepoint models of dynamic systems. *In preparation*

J. T. A. Koster (1996). Markov properties of nonrecursive causal models. *Annals of Statistics*, vol. 24, no. 5, pp. 2148–2177.

G. Lacerda, P. Spirtes, J. Ramsey, P. O. Hoyer (2008). Discovering cyclic causal models by independent components analysis. *Proceedings of the Twenty-Fourth Annual Conference on Uncertainty in Artificial Intelligence*.

U. Nodelman, C. Shelton, and D. Koller (2002). Continuous time Bayesian networks. *Eighteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 378–387.

U. Nodelman, C. Shelton, and D. Koller (2003). Learning continuous time Bayesian networks. *Nineteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 451–458.

J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kauffman.

J. Pearl (1995). Causal diagrams for empirical research. *Biometrika*, vol. 82, no. 4, pp. 669–710.

J. Pearl (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

J. Pearl and R. Dechter (1996). Identifying independence in causal graphs with feedback. *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 420–426.

L. Quinlan, S. Faherty, and M. Kane (2003). Phospholipase C and protein kinase C involvement in mouse embryonic stem-cell proliferation and apoptosis. 126(1):121-31. *Reproduction*

T. S. Richardson (1996). A discovery algorithm for directed cyclic graphs. *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 454–461.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan (2005). Causal protein-signaling networks derived from multiparameter single-cell data. 308(5721):523 - 529.*Science*.

S. Sharma, S., Durvasula S., Dillan, W. R.. Some results on the behavior of alternate covariance structure estimation in the presence of non-normal data. *Journal of Marketing research* 26, 214-221. (1989)

T. Shimizu, T. Kato, Jr. , A. Tachibana and M. S. Sasaki (1999). Coordinated Regulation of Radioadaptive Response by Protein Kinase C and p38 Mitogen-Activated Protein Kinase. 251(2):424-32. *Experimental Cell Research*

P. Spirtes, C. Glymour, and R. Scheines (1993). *Causation, Prediction and Search*, Lecture Notes in Statistics, vol. 81. Springer-Verlag.

P. Spirtes (1995). Directed cyclic graphical representations of feedback models. *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*, pp. 491–498.

A. Xu, Y. Wang, L. Yi Xu, and R. Stewart Gilmour (2001). Protein Kinase C-mediated Negative Feedback Regulation Is Responsible for the Termination of Insulin-like Growth Factor I-induced Activation of Nuclear Phospholipase C 1 in Swiss 3T3 Cells. 276(18):14980-6. *J. Biol. Chem*