Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Indirect Causes in Dynamic Bayesian Networks Revisited

Alexander Motzek and **Ralf Möller**
Institute of Information Systems
Universität zu Lübeck, Germany
{motzek,moeller}@ifis.uni-luebeck.de

Abstract

Modeling causal dependencies often demands cycles at a coarse-grained temporal scale. If Bayesian networks are to be used for modeling uncertainties, cycles are eliminated with dynamic Bayesian networks, spreading indirect dependencies over time and enforcing an infinitesimal resolution of time. Without a "causal design," i.e., without anticipating indirect influences appropriately in time, we argue that such networks return spurious results. By introducing activator random variables, we propose template fragments for modeling dynamic Bayesian networks under a causal use of time, anticipating indirect influences on a solid mathematical basis, obeying the laws of Bayesian networks.

1 Introduction

Dynamic Bayesian networks (DBNs) are an extension to Bayesian networks motivated from two perspectives, on the one hand as a manifestation of cyclic dependencies over time, closely related to Markov models [Murphy, 2002], on the other hand as a stationary process repeated over time in fixed time slices [Glesner and Koller, 1995]. Considering [Pearl, 2002] who emphasized that Bayesian networks should be a direct representation of the world instead of a reasoning process, both views seem to be conflicting. A stationary model repeated over time with cyclic dependencies would expand to infinity already for one timeslice. Therefore, cyclic dependencies in a stationary process are restricted [Jaeger, 2001] and forced into a strict order, e.g., state variables of time t are only dependent of states at $t-1$. Unfortunately, this means that evidence at a certain time point does not affect states at this time point, but one slice later.

In the extreme form of a DBN, every state variable is dependent on every other. In that case, there is no option to leave such dependencies in their causally correct same timestep as every dependency would cause cycles. Therefore, states can only be dependent on states from a previous timestep. However, this poses conflicts in causality, as a) the temporal causality is simply inaccurate and b) no indirect effects are considered, enforcing an infinitesimal resolution of time instead of a world-representing designed time and heavily limits the usage of a DBN.

To circumvent this problem, basically two options are available. As investigated by [Boutilier *et al.*, 1996] variables might be independent in certain contexts, which would allow a causally correct network generation from rules such as those presented in [Glesner and Koller, 1995] or [Ngo and Haddawy, 1997]. Still, then rules would need to be designed with a procedural view, degrading a BN to a procedural tool in a reasoning process, rather than designing it as a first-class declarative representation. Further, such rules would inherently be cyclic and might cause problems as stated by [Ngo *et al.*, 1995]. A second option would be to heavily restrict a DBN to specialized observation sets, e.g., to "single observations at a time" as done in [Sanghai *et al.*, 2005], s.t. no indirect causes need to be considered. Again, this implies that observations are made at an infinitesimal resolution of time.

The contribution of this paper can be summarized as follows. By introducing activator random variables, we propose template fragments for modeling DBNs under an unrestricted use of time, anticipating indirect influences on a solid mathematical basis, obeying the laws of Bayesian networks. This is beneficial for application contexts where causal models arise naturally and require a view over time, e.g., automatic learning of causal influences from coarse observation sets and—as a long-term goal—finding causally correct explanations and relations in (temporally uncertain) knowledge bases requiring anticipation of causal chains, e.g., DeepQA [Ferrucci *et al.*, 2010] or the Knowledge Vault [Murphy *et al.*, 2014].

We discuss preliminaries on DBNs and context-specific independencies as introduced by [Boutilier *et al.*, 1996] and [Haddawy *et al.*, 1995] in Sec. 2. By extending DBNs with activator random variables, we propose Activator Dynamic Bayesian Networks (ADBNs) in Sec. 3, derive common operations on ADBNs such as filtering and smoothing in Sec. 4, discuss our results in Sec. 5, and conclude with Sec. 6.

2 Dynamic Bayesian Networks: Preliminaries

A DBN models a stationary Markov process of state influences and transitions that is repeated over time.

Notation 2.1 (State Variables). *Let X_i^t be the random variable of the i^{th} state variable X_i at time t, where X_i^t is assignable to one of its possible values $x_i \in \mathrm{dom}(X_i^t)$. Let $\vec{X}^t$ be the set of all n state variables at time t, s.t.,*

$$\vec{X}^t = \left(X_1^t, \ldots, X_n^t\right)^\intercal .$$

Let $P(X_i^t = x_i)$ (or $P(x_i^t)$ for brevity) denote the probability of state X_i having x_i as a value at time t. If $\mathrm{dom}(X) = \{true, false\}$ we write x^t for the event $X^t = true$ and $\neg x^t$ for $X^t = false$ as usual. If X_i^t is unspecified and not defined through a query, $P(X_i^t)$ denotes the probability distribution of X_i^t for all its possible values.

Definition 1 (Dynamic Bayesian Network). A DBN is a tuple $(B_0, B_\rightarrow)$ with B_0 defining an initial Bayesian network (BN) representing time $t = 0$ containing all states X_i^0 in $\vec{X}^0$ and a consecutively repeated Bayesian network fragment $B_\rightarrow$ defining state dependencies between X_i^s and X_j^t, with $X_i^s \in \vec{X}^s, X_j^t \in \vec{X}^t, s \leq t$. By repeating $B_\rightarrow$ for every time step $t > 0$, a DBN $(B_0, B_\rightarrow)$ is unfolded into a BN uniquely defining a joint probability $P(\vec{X}^{0:t^\intercal})$. Notwithstanding, for every random variable X_i^t a local conditional probability distribution (CPD), e.g., as a CPT, is defined.

State dependencies defined in $B_\rightarrow$ are limited, s.t. no cyclic dependencies are created during unfolding. For $t-1 \leq s \leq t$, we speak of a first order Markov property, which we want to discuss in this paper. For any probabilistic model with $t-1 \leq s < t$, i.e., states at time t are only dependent of states at time $t-1$, an acyclicity constraint in the directed graph holds. A limited set of dependencies of the form $t-1 \leq s \leq t$ are possible, as long as no directed cycles are created. ▲

Commonly, in such networks we distinguish between observable (sensors) and unobservable (hidden states) variables. For our work, we consider a fully observable Markov model containing only observable *states*.

Diagonal state dependencies (as in Fig. 1) with $t-1 \leq s < t$ (acc. to Def. 1) are often due to syntactic constraints on (D)BNs and stand in conflict with an actual causality. Such dependencies exist causally at $s = t$, but create directed cycles. While conflicting with causality, further, dependencies on "sibling" states of one time slice are spread over the past. This means, indirect causes among siblings are not anticipated, or rather, that chain reactions are not covered.

Example 2.1 (Regulatory Compliance). *In a company deliberately placed false information, e.g., faked payment sums for bribe money, might divulge throughout a company until every employee believes (unknowingly) in a lie. We therefore model a probabilistic regulatory-compliance checking tool using a DBN to track and query possible violations of employees over time and to track back potential sources of deliberately placed false information. If one employee believes in false information, we do not say that such an employee is "corrupt", but say that he is* credulous*. Every employee,* Claire, Don *and* Earl, *say, is represented by one state in $\vec{X}^t$. The probability $P(X_i^t)$, encodes the belief in employee X_i being credulous x_i^t or being integrous $\neg x_i^t$ at time t. We model B_0, s.t., it models our prior belief in every employee being a source of false information, i.e., B_0 is a BN containing all $\vec{X}^0$ as prior random variables; say $P(c^0) = 0.9, P(d^0) = 0.6, P(e^0) = 0.01$. Being credulous is permanent, such that $B_\rightarrow$ describes all random variables X_i^t depending on X_i^{t-1} with conditional probability $P(x_i^t | x_i^{t-1}) = 1$.*

An employee might influence another employee in his writings or, rather, in his information state. A credulous employee might therefore (undeliberately) influence his colleague such that the colleague also believes in false information, i.e., becomes credulous, too. Say, Claire influences Don, and if Claire is credulous, there is a probability of Don becoming credulous too. Further, if Don influences Earl, there is a probability that Claire influences Earl indirectly *through Don. We can model this correctly as a dependency as $C^t \rightarrow D^t \rightarrow E^t$ in $B_\rightarrow$. We assume an individual probability of 0.8 for an employee becoming credulous when being influenced by a credulous person and a noisy-or combination for every state.*

However, we want to model that all employees can influence each other, and, to assure an acyclicity constraint, we must bend the influencing dependency to a consecutive timestep in $B_\rightarrow$ (as in Fig. 1). This is unavoidable, but is inaccurate from a world representation point of view, as indirect influences are now anticipated spuriously. Earl is now influenced by Claire through *Don from a Claire of the penultimate time. This means, a time slice must be infinitesimal small, e.g. secondly, to anticipate all indirect influences, and cannot be chosen freely to an intended use case, e.g. daily.*

In our example, we now can make observations, e.g., from unheralded compliance checkups and trace a potential diffusion of false information throughout our company over time. Still, we cannot actually model an accurate representation of a world, because we have to use a modeled dimension (time) for assuring syntactic constraints of (in)dependencies.

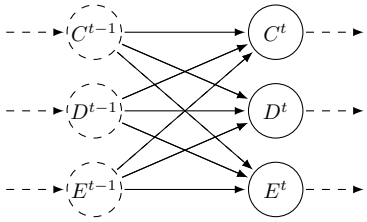

Figure 1: A "diagonal" DBN fragment $B_\rightarrow$ for Ex. 2.1. E^t is only influenced indirectly by a past C^{t-2} through D^{t-1}.

Classically a conditional independency in a Bayesian network is represented by the lack of an arc between two nodes. Another kind of independencies in Bayesian networks, called context-specific independencies (CSIs), has been studied by [Boutilier *et al.*, 1996] & [Ngo and Haddawy, 1997] and has mainly been used for more efficient inference in such networks. CSIs represent dependencies in a BN that are only present in specific contexts. We extend this idea by defining special activator random variables.

Definition 2 (Activator Random Variable). We define A_{XY} to be an activator random variable which activates a dependency of random variable Y on X in a given context. Let $\mathrm{dom}(A_{XY}) = \{true, false\}$ (extensions to non-boolean domains are straightforward). We define the *deactivation* criterion from a functional perspective towards the CPT as

$$\forall x, x' \in \mathrm{dom}(X), \forall y \in \mathrm{dom}(Y), \forall \vec{z} \in \mathrm{dom}(\vec{Z}): \quad (1)$$
$$P(y|x, \neg a_{XY}, \vec{z}) = P(y|x', \neg a_{XY}, \vec{z}) = P(y|*, \neg a_{XY}, \vec{z}),$$

where $*$ represents a wildcard and $\vec{z}$ further dependencies.

The *activation* criterion describes a situation where Y becomes dependent on X, i.e., the CPT entry for y is not uniquely identified by just a_{XY} and $\vec{z}$, hence

$$\exists x, x^* \in \mathrm{dom}(X), \exists y \in \mathrm{dom}(Y), \exists \vec{z} \in \mathrm{dom}(\vec{Z}) : P(y|x, a_{XY}, \vec{z}) \neq P(y|x^*, a_{XY}, \vec{z}) \; . \quad \blacktriangle \qquad (2)$$

Example 2.2 (Activator). *Claire does not constantly influence Don, but only if Claire sends a letter to Don. We can observe possible exchanges from used envelopes (possibly found in the trash bin). On such envelopes, we find multiple transfers from a coarse time interval in an imprecise or inaccurate order. For example, a transfer from Don to Earl and one from Claire to Don might include a transitive influence of Claire on Earl* at the same time. *A document transfer at time t, denoted M^t_{CD}, is then an activator for an influence of Claire C^t on Don D^t. Likewise, if we can neglect this document transfer, i.e. observe $\neg m^t_{CD}$, Don becomes independent of Claire at time t.*

The example shows that sometimes dependencies are modeled in $B_{\rightarrow}$ that are not always needed.

3 Activator Dynamic Bayesian Networks

We extend Bayesian networks such that, besides state variables, we have activators purely acting as necessary conditions for context-specific (in)dependencies.

Notation 3.1 (Activator Matrices). *Let $A^{s\,t}$ describe the matrix of all activator random variables between time-slice s and t, s.t.,*

$$A^{s\,t} = \begin{pmatrix} A^{s\,t}_{11} & \cdots & A^{s\,t}_{1n} \\ \vdots & \ddots & \vdots \\ A^{t}_{n1} & \cdots & A^{s\,t}_{nn} \end{pmatrix} \; .$$

Let $\vec{A}^{s\,t}_i$ denote the i^{th} column of $A^{s\,t}$ and let $\vec{\mathcal{A}}^{s\,t}$ denote the corresponding column vector of all entries of $A^{s\,t}$, s.t.

$$\vec{\mathcal{A}}^{s\,t} = \left(A^{s\,t}_{11}, \ldots, A^{s\,t}_{1n}, \ldots, A^{s\,t}_{n1}, \ldots, A^{s\,t}_{nn}\right)^{\intercal} \; .$$

Definition 3 (Activator Dynamic Bayesian Network (ADBN)). An ADBN fragment template $B'_{\rightarrow}$ consists of dependencies between states X^s_i and X^t_j, $t-1 \leq s \leq t$ (Markov-1) and matrices $A^{s\,t}$ of activators. Let $A^{s\,t}_{ij}$ be the activator random variable influencing X^t_j regarding a dependency on X^s_i, such that X^t_j's local CPT follows Eq. 2 and Eq. 1. Every activator is assigned a prior probability. An ADBN is then syntactically defined by $(B_0, B'_{\rightarrow})$ defining its semantic as a joint probability $P(\vec{X}^{0:t\intercal}, \vec{\mathcal{A}}^{01:tt\intercal})$. $\blacktriangle$

Still, $t-1 \leq s < t$ is necessary to absolutely assure an acyclicity constraint when modeling ADBNs. Under the condition $t-1 \leq s \leq t$, possibly directed cycles are created. For our work we only consider the problematic $t-1 \leq s \leq t$ case, only containing in-time-slice dependencies (as in Fig. 2). For brevity, we write A^t for A^{tt} excluding A^{tt}_{kk}, and, correspondingly, $\mathcal{A}^t$ and A^t_{ij}.

The constraint $t-1 \leq s < t$ for assuring acyclicity even in complete digraphs limits a causal reasoning process to direct dependencies between states.

Proposition 1 (Diagonal (A)DBN Restrictions). A classic, "diagonal" (as in Fig. 2) (A)DBN of type $t-1 \leq s < t$ is restricted in its usage to special observation sets. Indirect influences are spread over multiple timesteps and possible indirect influences inside one timestep cannot be considered. This enforces a) an infinitesimal resolution of observations, where indirect effects do not need to be anticipated or b) restricts a DBN to observations where indirect influences strictly do not occur. This implies, not a single two activators A^t_{*i} and A^t_{i*} are allowed to be probably active, i.e. the set of probably active activators must form a bipartite digraph with uniformly directed edges. Further, only up to $n^2/4$ activators are allowed to be probably active per timestep, and *all* other activators must be (i) observed to be (ii) deactive. If, in a diagonal DBN, observations can neither hold a) or b), observation- and query-(de)serializations would be needed, and $n-2$ spurious "time"-slices would need to be inserted between $t-1$ and t. In our opinion, this degrades a BN to a reasoning tool. $\blacktriangle$

We show, by using a modified acyclicity constraint, that in ADBNs (Fig. 2) we can correctly anticipate indirect influences by modeling dependencies causally correctly.

Example 3.1 (Example continued). *By extending our credulousness representing DBN with document transfer activators we obtain an ADBN with activators $\vec{\mathcal{A}}^t = (M^t_{CD}, M^t_{DC}, M^t_{DE}, M^t_{ED}, M^t_{CE}, M^t_{EC})$ and states $\vec{X}^t = (C^t, D^t, E^t)$. For every t we assume a prior probability for any transfer of $P(m^t_{ij}) = 0.5$. Still, we have to assume that every employee can influence every other, i.e., send him a document. To cover chain reactions of multiple transfers, we would then need syntactically forbidden cyclic dependencies.*

The following Thm. 1 states that by using an ADBN it is indeed possible to move an acyclicity constraint from a design phase to an operations phase while maintaining a solid mathematical basis in accordance with Bayesian network semantics. This means, if the only possible mails are from Claire to Don to Earl, we could have modeled all influences correctly in one timestep during design of the DBN. Unfortunately, we do not know possible observations during design and such acyclic mail exchanges may differ in every time step.

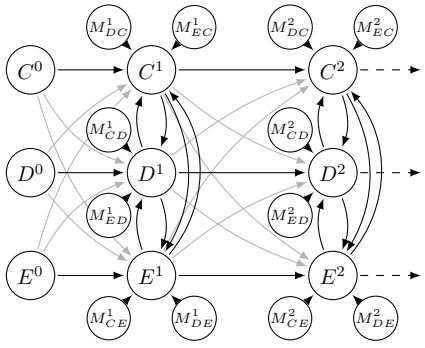

Figure 2: A causally correctly represented world using an ADBN for Ex. 3.1. Syntactic DAG constraints of (D)BNs prevented desired cyclic dependencies in this design and "diagonal" state dependencies were enforced (hinted in light grey). In the diagonal case, M^t_{XY} represents $M^{t-1\,t}_{XY}$, i.e. M^t_{XY} affects the dependency of state Y^t on X^{t-1}.

Notation 3.2 (Vector Operands). *Let* $\mathbf{f}_\Gamma(\vec{X}, Y)$ *be the product of applied operands* f *to every row* $\{1 \leq i \leq \text{rank}(\vec{X})\}\backslash\Gamma$*, i.e., we iterate over every row of* $\vec{X}$ *without rows in the set* Γ *and apply* f *to this row's elements. Scalars* Y *are used in every row, i.e.,*

$$\mathbf{f}_\Gamma(\vec{X}, Y) = \prod_{i\backslash\Gamma} f(X_i, Y)$$

Notation 3.3 (Lexicographic Order). *Let* $\prec$ *be a lexicographic term order, such that* $X_*^{t-1} \prec X_*^t$*,* $X_i^t \prec X_{i+1}^t$*, and* $A_{**}^{t-1} \prec A_{**}^t$*,* $A_{i*}^t \prec A_{(i+1)*}^t$*,* $A_{ij}^t \prec A_{i(j+1)}^t$*, and* $A_{**}^t \prec X_*^t$*,* $X_*^{t-1} \prec A_{**}^t$*.*

Theorem 1 (Bayesian Network Soundness). For every set of combinations of $\vec{\mathcal{A}}^{1:t}$ an ADBN (as in Fig. 2) corresponds to a Bayesian network, if, for all t, $\vec{\mathcal{A}}^t$ satisfies the new acyclicity constraint:

$$\begin{aligned} \forall x, y, z \in \vec{X}^t : \mathscr{A}(x,z)^t, \mathscr{A}(z,y)^t \rightarrow \mathscr{A}(x,y)^t \\ \neg\exists q : \mathscr{A}(q,q)^t \,, \end{aligned} \tag{3}$$

with a function $\mathscr{A}(i,j)^t$ that is defined as

$$\mathscr{A}(i,j)^t = \begin{cases} \mathit{false} & \text{if} \quad A_{ij}^t = \neg a_{ij}^t \\ \mathit{true} & \text{if} \quad \text{else} \end{cases} .$$

Following the lexicographic order, the joint probability (JP) $P(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal})$ of an ADBN is specified by,

$$\begin{aligned} &P(X_1^0) \cdot \ldots \cdot P(X_n^0) \\ &\cdot \prod_{i=1}^{t} P(X_1^i | X_2^i, \ldots, X_n^i, A_{21}^i, \ldots, A_{n1}^i, X_1^{i-1}) \cdot \ldots \\ &\cdot P(X_n^i | X_1^i, \ldots, X_{n-1}^i, A_{1n}^i, \ldots, A_{(n-1)n}^i, X_n^{i-1}) \\ &\qquad \cdot P(A_{12}^i) \cdot \ldots \cdot P(A_{n(n-1)}^i) \,, \end{aligned}$$

written for brevity using Not. 3.2 as

$$\mathbf{P}(\vec{X}^0) \;\cdot\; \prod_{i=1}^{t} \mathbf{P}(\vec{X}^i | \vec{X}^{i^\intercal} \backslash \vec{X}^i, A^{\intercal i}, \vec{X}^{i-1}) \;\cdot\; \mathbf{P}(\vec{\mathcal{A}}^i) \quad .$$

As expected, the JP can be defined recursively:

$$\begin{aligned} P(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal}) &= P(\vec{X}^{0:t-1^\intercal}, \vec{\mathcal{A}}^{1:t-1^\intercal}) \\ &\cdot \mathbf{P}(\vec{X}^t | \vec{X}^{t^\intercal} \backslash \vec{X}^t, A^{\intercal t}, \vec{X}^{t-1}) \cdot \mathbf{P}(\vec{\mathcal{A}}^t) \,. \quad \blacktriangle \end{aligned} \tag{4}$$

Informally, Eq. 3 states that a deactive activator must break open dependency cycles, i.e., the set of possibly active activators forms a directed acyclic graph (DAG).

Proof of Theorem 1. We show that for every set of combinations of $\vec{\mathcal{A}}^{1:t}$ the joint probability stated in Thm. 1 is unique and well-defined, iff for all t the set of $\vec{\mathcal{A}}^t$ obeys Eq. 3. We show this by reversing conditional independency assumptions in the semantic JP and find the unique topological order of our syntactical graph structure.

We begin with B_0, which can be written as

$$\begin{aligned} P(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal}) &= P(X_1^0) \cdot \ldots \cdot P(X_n^0) \cdot \gamma \\ &= P(X_1^0, \ldots, X_n^0) \cdot \gamma = P(\vec{X}^{0^\intercal}) \cdot \gamma \,, \end{aligned} \tag{5}$$

with

$$\gamma = \prod_{i=1}^{t} \mathbf{P}(\vec{X}^i | \vec{X}^{i^\intercal} \backslash \vec{X}^i, A^{\intercal i}, \vec{X}^{i-1}) \cdot \mathbf{P}(\vec{\mathcal{A}}^i) \,.$$

Consecutively, we roll up the joint distribution according to Bayes' chain rule. Considering an extreme case of a set of activators corresponding to Eq. 3, it is straightforward that under Eq. 3 there must always $\exists X_{E1}^1 : \forall i A_{i(E1)}^1 = \neg a_{i(E1)}^1$, such that due to Eq. 1, the set of activators and previous states uniquely identify the CPT entry and X_{E1}^1 becomes independent of all other $\vec{X}^1$, such that the JP can be written as

$$\begin{aligned} &P(\vec{X}^{0^\intercal}) \cdot P(X_{E1}^1 | *, \vec{A}_{E1}^{1^\intercal}, X_{E1}^0) \\ &\quad \cdot \mathbf{P}_{\{E1\}}(\vec{X}^1 | \vec{X}^{1^\intercal} \backslash \vec{X}^1, A^{\intercal 1}, \vec{X}^0) \cdot \mathbf{P}(\vec{\mathcal{A}}^1) \\ &\quad \cdot \prod_{i=2}^{t} \mathbf{P}(\vec{X}^i | \vec{X}^{i^\intercal} \backslash \vec{X}^i, A^{i^\intercal}, \vec{X}^{i-1}) \cdot \mathbf{P}(\vec{\mathcal{A}}^i) \,. \end{aligned} \tag{6}$$

By reversing X_{E1}^1's conditional independency we can write

$$\begin{aligned} &P(\vec{X}^{0^\intercal}) \cdot P(X_{E1}^1 | *, \vec{\mathcal{A}}^{1^\intercal}, \vec{X}^{0^\intercal}) \cdot \mathbf{P}(\vec{\mathcal{A}}^1) \\ &\quad \cdot \mathbf{P}_{\{E1\}}(\vec{X}^1 | \vec{X}^{1^\intercal} \backslash \vec{X}^1, A^{\intercal 1}, \vec{X}^0) \\ &\quad \cdot \prod_{i=2}^{t} \mathbf{P}(\vec{X}^i | \vec{X}^{i^\intercal} \backslash \vec{X}^i, A^{\intercal i}, \vec{X}^{i-1}) \cdot \mathbf{P}(\vec{\mathcal{A}}^i) \,. \end{aligned}$$

Hence, with

$$\begin{aligned} \mathbf{P}(\vec{\mathcal{A}}^t) &= P(A_{12}^t) \cdot \ldots \cdot P(A_{1n}^t) \cdot \ldots \cdot P(A_{n1}^t) \cdot \ldots \cdot P(A_{n(n-1)}^t) \\ &= P(A_{12}^t, \ldots, A_{1n}^t, \ldots, A_{n1}^t, \ldots, A_{n(n-1)}^t) \\ &= P(\vec{\mathcal{A}}^{t^\intercal}) \,, \end{aligned}$$

we can combine $P(\vec{X}^{0^\intercal})$ with $P(\vec{\mathcal{A}}^{1^\intercal})$ to $P(\vec{\mathcal{A}}^{1^\intercal}, \vec{X}^{0^\intercal})$, s.t. the first eliminated state variable X_{E1}^1 can be combined to

$$\begin{aligned} &P(X_{E1}^1, \vec{\mathcal{A}}^{1^\intercal}, \vec{X}^{0^\intercal}) \\ &\quad \cdot \mathbf{P}_{\{E1\}}(\vec{X}^1 | \vec{X}^{1^\intercal} \backslash \vec{X}^1, A^{\intercal 1}, \vec{X}^0) \\ &\quad \cdot \prod_{i=2}^{t} \mathbf{P}(\vec{X}^i | \vec{X}^{i^\intercal} \backslash \vec{X}^i, A^{\intercal i}, \vec{X}^{i-1}) \cdot \mathbf{P}(\vec{\mathcal{A}}^i) \,. \end{aligned}$$

Consecutively, there $\exists X_{E2}^1 : \forall \{i \backslash E1\}\, A_{i(E2)}^1 = \neg a_{i(E2)}^1$, s.t.,

$$\begin{aligned} &P(X_{E1}^1, \vec{\mathcal{A}}^{1^\intercal}, \vec{X}^{0^\intercal}) \cdot P(X_{E2}^1 | *, X_{E1}^1, *, \vec{A}_{E2}^{1^\intercal}, X_{E2}^0) \\ &\quad \cdot \mathbf{P}_{\{E1,E2\}}(\vec{X}^1 | \vec{X}^{1^\intercal} \backslash \vec{X}^1, A^{\intercal 1}, \vec{X}^0) \\ &\quad \cdot \prod_{i=2}^{t} \mathbf{P}(\vec{X}^i | \vec{X}^{i^\intercal} \backslash \vec{X}^i, A^{\intercal i}, \vec{X}^{i-1}) \cdot \mathbf{P}(\vec{\mathcal{A}}^i) \,, \end{aligned}$$

for which we can reverse the conditional independency again and obtain

$$\begin{aligned} &P(X_{E1}^1, \vec{\mathcal{A}}^{1^\intercal}, \vec{X}^{0^\intercal}) \cdot P(X_{E2}^1 | *, X_{E1}^1, *, \vec{\mathcal{A}}^{1^\intercal}, \vec{X}^{0^\intercal}) \\ &\quad \cdot \mathbf{P}_{\{E1,E2\}}(\vec{X}^1 | \vec{X}^{1^\intercal} \backslash \vec{X}^1, A^{\intercal 1}, \vec{X}^0) \\ &\quad \cdot \prod_{i=2}^{t} \mathbf{P}(\vec{X}^i | \vec{X}^{i^\intercal} \backslash \vec{X}^i, A^{\intercal i}, \vec{X}^{i-1}) \cdot \mathbf{P}(\vec{\mathcal{A}}^i) \,, \end{aligned}$$

which, according to Bayes' chain rule, can be written as

$$P(X^1_{E2}, X^1_{E1}, \vec{\mathcal{A}}^{1\intercal}, \vec{X}^{0\intercal}) \cdot \mathbf{P}_{\{E1,E2\}}(\vec{X}^1|\vec{X}^{1\intercal}\backslash\vec{X}^1, A^{\intercal 1}, \vec{X}^0) \cdot \prod_{i=2}^{t} \mathbf{P}(\vec{X}^i|\vec{X}^{i\intercal}\backslash\vec{X}^i, A^{\intercal i}, \vec{X}^{i-1}) \cdot \mathbf{P}(\vec{\mathcal{A}}^i)\,.$$

Consecutively repeating this process for every further X_{Ei}, where the i^{th} elimination variable is maximally dependent on the previous $(i-1)$ elimination variables, we, henceforth, approach the elimination of X^1_{En}, which is dependent on up to every other $\vec{X}^1$, which are in fact all eliminated variables up to now, s.t.

$$P(X^1_{E(n-1)}, \ldots, X^1_{E1}, \vec{\mathcal{A}}^{1\intercal}, \vec{X}^{0\intercal}) \cdot P(X^1_{En}|X^1_{E(n-1)}, \ldots, X^1_{E1}, \vec{A}^{1\intercal}_{En}, X^0_{En}) \cdot \prod_{i=2}^{t} \mathbf{P}(\vec{X}^i|\vec{X}^{i\intercal}\backslash\vec{X}^i, A^{\intercal i}, \vec{X}^{i-1}) \cdot \mathbf{P}(\vec{\mathcal{A}}^i)\,,$$

for which we can reverse the conditional independency again and combine the JP finally to

$$P(X^1_{En}, X^1_{E(n-1)}, \ldots, X^1_{E1}, \vec{\mathcal{A}}^{1\intercal}, \vec{X}^{0\intercal}) \cdot \prod_{i=2}^{t} \mathbf{P}(\vec{X}^i|\vec{X}^{i\intercal}\backslash\vec{X}^i, A^{\intercal i}, \vec{X}^{i-1}) \cdot \mathbf{P}(\vec{\mathcal{A}}^i)\,.$$

Indeed, we already obtained a partial topological order $>$ of $X^1_{En} > X^1_{E(n-1)} > \ldots > X^1_{E1} > \vec{\mathcal{A}}^{1\intercal} > \vec{X}^{0\intercal}$. Following this procedure for the remaining t, we finally obtain

$$P(X^t_{E(n-1)}, ..., X^t_{E1}, \vec{\mathcal{A}}^{t\intercal}, ..., X^1_{En}, ..., X^1_{E1}, \vec{\mathcal{A}}^{1\intercal}, \vec{X}^{0\intercal}) \cdot P(X^t_{En}|X^t_{E(n-1)}, \ldots, X^1_{E1}, A^{1\intercal}_{En}, X^{t-1}_{En})\,.$$

With a final reverse conditional independency assumption,

$$P(X^t_{E(n-1)}, ..., X^t_{E1}, \vec{\mathcal{A}}^{t\intercal}, ..., X^1_{En}, ..., X^1_{E1}, \vec{\mathcal{A}}^{1\intercal}, \vec{X}^{0\intercal}) \cdot P(X^t_{En}|X^t_{E(n-1)}, ..., X^t_{E1}, \vec{\mathcal{A}}^{t\intercal}, ..., X^1_{En}, ..., X^1_{E1}, \vec{\mathcal{A}}^{1\intercal}, \vec{X}^{0\intercal})\,,$$

we obtain a complete topological order and a unique JP of

$$P(X^t_{En}, X^t_{E(n-1)}, ..., X^t_{E1}, \vec{\mathcal{A}}^{t\intercal}, ..., X^1_{En}, ..., X^1_{E1}, \vec{\mathcal{A}}^{1\intercal}, \vec{X}^{0\intercal})\,. \tag{7}$$

We have shown that the claimed JP of Th. 1 in fact is a unique and well-defined JP defining a topological order of a corresponding Bayesian network. □

Informally speaking, this proof shows that in an ADBN an acyclicity constraint can be postponed to an operation phase while assuring soundness with BNs. This means, a BN can actually be a *cyclic* graph, if it is used correctly.

Proposition 2 (Completeness). An ADBN can model any JP. We have shown that in an ADBN all random variables of t can be dependent on each other, as long as during operations only certain combinations of activators are used. This means, that any form of in-time-slice structure can be modeled through adequate specifications of $\vec{\mathcal{A}}^t$. Straightforwardly, this can, if causally needed, be extended to further "diagonal" dependencies between states of consecutive time slices, including activator random variables for those dependencies. Obviously, this does not create cyclic dependencies and thus satisfies Eq. 3, i.e., is an ADBN. Such an ADBN would contain all possible dependencies between states and leave the option to (de)activate them, meaning, represents the most general form of an Markov-1 ADBN template including all possible Markov-1 DBN structures. Noteworthy, we can directly embed [Pearl, 2009]'s do-calculus here using activators. ▲

4 Operations

Based on Thm. 1 marginalization is well-defined and filtering, smoothing and prediction (according to the meaning of [Murphy, 2002]) can be derived from the joint distribution. We derive those operations while carefully handling novel in-time-slice dependencies and activator random variables.

Notation 4.1 (Notation for Observations). *Let $\vec{Z}^t \subseteq \vec{X}^t$ be a set of observed and $\vec{\zeta}^t = \vec{X}^t\backslash\vec{Z}^t$ be the corresponding set of not-observed state variables. Let $B^t \subseteq A^t$ be a set of observed activators and $\vec{\mathcal{B}}^t \subseteq \vec{\mathcal{A}}^t$ be the corresponding column vector representation. Likewise, let $\vec{\beta}^t = \vec{\mathcal{A}}^t\backslash\vec{\mathcal{B}}^t$ be the column vector of all not-observed activators. We write $\vec{z}^t$ for $\vec{Z}^t = \vec{z}$ and $\vec{b}^t$ for $\vec{\mathcal{B}}^t = \vec{b}$.*

Every query contradicting with observations, i.e., $\vec{x}^t$ contradicts $\vec{z}^t$, is defined to be of probability 0. Further, every observation in $\vec{z}^t$ uniquely defines its corresponding random variable in $\vec{X}^t$.

4.1 Filtering

We calculate the complete conditional joint probability $P(\vec{X}^{0:t\intercal}, \vec{\mathcal{A}}^{1:t\intercal}|\vec{z}^{0:t\intercal}, \vec{b}^{1:t\intercal})$ at every timestep, from which every desired filtering operation can be marginalized out. With a normalization factor α, filtering is generally defined from the JP as

$$P(\vec{X}^{t\intercal}, \vec{\mathcal{A}}^{t\intercal}|\vec{z}^{0:t\intercal}, \vec{b}^{1:t\intercal}) = \alpha \sum_{\vec{\zeta}^{0:t-1\intercal}} \sum_{\vec{\beta}^{1:t-1\intercal}} P(\vec{X}^{0:t\intercal}, \vec{\mathcal{A}}^{1:t\intercal})\,,$$

where all values of variables $\vec{X}^t, \vec{\mathcal{A}}^t$ are defined by the query and variables $\vec{X}^{0:t-1}, \vec{\mathcal{A}}^{1:t-1}$ are defined by either observations in the sets $\vec{z}^{0:t-1}, \vec{b}^{1:t-1}$ or through summation over unobserved variables in $\vec{\zeta}^{0:t-1}, \vec{\beta}^{1:t-1}$.

Definition 4 (Filtering). Using the recursive definition of the joint probability in Eq. 4, filtering is efficiently defined as

$$P(\vec{X}^{t\intercal}, \vec{\mathcal{A}}^{t\intercal}|\vec{z}^{0:t\intercal}, \vec{b}^{1:t\intercal}) = \alpha \cdot \sum_{\vec{\zeta}^{t-1\intercal}} \sum_{\vec{\beta}^{t-1\intercal}} P(\vec{X}^{t-1\intercal}, \vec{\mathcal{A}}^{t-1\intercal}|\vec{z}^{0:t-1\intercal}, \vec{b}^{1:t-1\intercal}) \cdot \mathbf{P}(\vec{X}^t|\vec{X}^{t\intercal}\backslash\vec{X}^t, A^{\intercal t}, \vec{X}^{t-1}) \cdot \mathbf{P}(\vec{\mathcal{A}}^t)\,. \quad ▲ \tag{8}$$

ADBN filtering from $t-1$ to t has time and space complexity $\mathcal{O}(1)$. Every incremental ADBN filtering is n-times faster than performing it in a serialized fashion having further $\mathcal{O}(t)$ space complexity for storing all orders. Further effort would be needed for generating such a serialized order.

Example 4.1 (Filtering). *With Thm. 1 we can actually model cyclic dependencies as desired in Ex. 3.1 and build an ADBN for our example as shown in Fig. 2.*

Say, Don and Earl did pass an initial checkup, but Claire did not. At $t = 1$, we observe a document transfer from Claire to Don, we are unsure about one from Don to Earl, but can neglect all other transfers. As Claire is credulous, we expect her to influence Earl slightly through *Don, expressible in the filtering operation $P(E^1|\vec{z}^{0:1\intercal}, \vec{b}^{1\intercal})$, with $\vec{z}^{0:1} = (c^0, \neg d^0, \neg e^0)^\intercal$, and $\vec{b}^1 = (m^1_{CD}, \neg m^1_{DC}, \neg m^1_{ED}, \neg m^1_{CE}, \neg m^1_{EC})^\intercal$.*

A diagonal DBN cannot anticipate the indirect influence, because t_1-Earl is influenced by a t_0-Don that has not received a document from Claire. This means our belief in Earl remains at 0 due to our initial observation of $\neg e^0$, i.e. $P'(E^1|\vec{z}^{0:1\intercal}, \vec{b}^{1\intercal}) = \langle 0, 1\rangle$. As $\vec{b}^1$ fulfills Eq. 3 we correctly obtain $P(E^1|\vec{z}^{0:1\intercal}, \vec{b}^{1\intercal}) = \langle 0.32, 0.68\rangle$ using an ADBN, i.e. we anticipate that Earl is influenced by Claire through Don.

To achieve the same result in a diagonal DBN, we need observations at a finer time scale, where all indirect influences are serialized, e.g., we must first observe m^1_{CD}, anticipated in the filtering operation $P'(E^1|\vec{z}^{0:1\intercal}, \vec{b}^{1\intercal})$ and then insert a "correcting" "time"-slice $t = 1.1$, where we anticipate the possibilities of M^1_{DE} in another operation $P'(E^{1.1}|\vec{z}^{0:1.1\intercal}, \vec{b}^{1:1.1\intercal})$. To achieve the result of one ADBN operation, we need $n - 1$ "diagonal"-operations.

Prediction is a filtering operation with an empty observation set. However, as a minimal set of observations is needed to remove cycles, plain prediction is not possible in our syntax. However, by splitting a prediction-observation-set into two subsets, e.g. first the lower triangle and second the upper triangle of A^{t+1} is observed to be deactive, prediction becomes possible. While this does not cover all possible chain reactions, it then covers significantly more than a classic DBN could cover (none), as previously discussed in Sec. 3 and Prop. 1.

4.2 Smoothing

Similar to the filtering operation, the general smoothing operation $P(\vec{X}^{k\intercal}, \vec{\mathcal{A}}^{k\intercal}|\vec{z}^{0:t\intercal}, \vec{b}^{1:t\intercal})$, $k < t$ can be derived from the joint probability as

$$P(\vec{X}^{k\intercal}, \vec{\mathcal{A}}^{k\intercal}|\vec{z}^{0:t\intercal}, \vec{b}^{1:t\intercal}) = \alpha \cdot \sum_{\vec{\zeta}^{0:k-1\intercal}} \sum_{\vec{\beta}^{1:k-1\intercal}} \sum_{\vec{\zeta}^{k+1:t\intercal}} \sum_{\vec{\beta}^{k+1:t\intercal}} P(\vec{X}^{0:t\intercal}, \vec{\mathcal{A}}^{1:t\intercal}) =$$

$$P(\vec{X}^{k\intercal}, \vec{\mathcal{A}}^{k\intercal}|\vec{z}^{0:k\intercal}, \vec{b}^{1:k\intercal}) \cdot P(\vec{z}^{k+1:t\intercal}, \vec{b}^{k+1:t\intercal}|\vec{X}^{k\intercal}, \vec{\mathcal{A}}^{k\intercal}),$$

in which we find a previous (stored) filtering operation, known as a forward message, and a new latter term commonly known in smoothing operations. Using an adequate recursive definition for the latter term, we obtain an efficient calculation method using a "backward message." The commonly known "sensor model" is, due to in-time-slice dependencies, included in the forward, as well as backward message.

Definition 5 (Smoothing). Smoothing at timestep k considering all evidences up to t is defined as

$$P(\vec{X}^{k\intercal}, \vec{\mathcal{A}}^{k\intercal}|\vec{z}^{0:t\intercal}, \vec{b}^{1:t\intercal}) = \alpha \cdot P(\vec{X}^{k\intercal}, \vec{\mathcal{A}}^{k\intercal}|\vec{z}^{0:k\intercal}, \vec{b}^{1:k\intercal}) \cdot \sum_{\vec{\zeta}^{k+1\intercal}} \sum_{\vec{\beta}^{k+1\intercal}} \mathbf{P}(\vec{X}^{k+1}|\vec{X}^{k+1\intercal} \backslash \vec{X}^k, A^{\intercal k+1}, \vec{X}^k) \cdot \mathbf{P}(\vec{\mathcal{A}}^{k+1}) \cdot P(\vec{z}^{k+2:t\intercal}, \vec{b}^{k+2:t\intercal}|\vec{X}^{k+1\intercal}, \vec{\mathcal{A}}^{k+1\intercal}). \quad (9)$$

The last term corresponds to the backward message and was calculated in the previous (i.e., previously calculated, but temporally consecutive) smoothing operation. ▲

Performing smoothing over all $k < t$ has $\mathcal{O}(t^2)$ time and constant space complexity or, by storing filtering operations, $\mathcal{O}(t)$ time and space complexity. Compared to a serialized version, without actually serializing, n^2-times faster or n-times faster and smaller.

Example 4.2 (Explaining away). *Continuing Ex. 4.1 this example demonstrates that smoothing handles explaining away over multiple timesteps and respects indirect causes. Say, only Don underwent a successful compliance check at time $t = 0$, i.e., $\vec{z}^0 = (\neg d^0)$. For $t = 1$ we found the same document transfer as previously, and for $t = 2$, a Sunday, we can neglect all, i.e., $\vec{b}^2 = \emptyset$. On that Sunday also irregularities in Earl's documents were found, i.e., $\vec{z}^2 = (e^2)$.*

If we perform the smoothing operation for Claire's initial belief state without considering evidence from $t = 2$, we end up with our prior belief of $P(C^0|\vec{z}^{0:1\intercal}, \vec{b}^{1\intercal}) = \langle 0.5, 0.5\rangle$, as we have gained no new information. However, with observations from $t = 2$, we need to consider an indirect influence by Claire onto Earl and our belief in her rises to $P(C^0|\vec{z}^{0:2\intercal}, \vec{b}^{1:2\intercal}) \approx \langle 0.532, 0.468\rangle$.

The slow increase is due to our high prior belief in Earl manipulating documents of $P(e^0) = 0.7$ and it is more likely that Earl has been manipulating documents ever since. If, say, Earl can be relieved from initial incriminations, i.e., $\neg e^0$, the only explanation for this situation is an indirect cause of Claire being credulous, which is correctly handled as $P(c^1, m^1_{DE}|\vec{z}^{0:2\intercal}, \vec{b}^{1:2\intercal}) = 1$. We can update our initial prior beliefs using smoothing and find that $P(d^0) = P(e^0) = 0$ but $P(c^0) = 1$. We can now say for sure, Claire is corrupt.

In a classic diagonal (A)DBN the last example would have been unexplainable, as indirect influences of t_1 (causally) would first be anticipated a step later at t_2 (for n=3). The detailed explanation is confusing, because it is not causal: at t_2, the time of incriminating evidence for Earl, we know that Earl is only influenced by himself, i.e. only t_1-Earl can be the source of his credulousness. At t_1, Earl only receives a document from integrous t_0-Don (observation). This is where the problem lies, t_0-Claire should have influenced t_0-Don by now, but t_0-Claire influences t_1-Don with her message m^1_{CD}. I.e. Earl cannot become credulous and the observation e^2 remains unexplainable. Mathematically we obtain $P(e^0|\vec{z}^{0:2\intercal}, \vec{b}^{1:2\intercal}) = 0$ because all terms in this calculation involve either the CPT entry $P(e^2|\underline{\neg m^2_{*E}}, C^1, D^1, \underline{\neg e^1}) = 0$ or $P(e^1|M^1_{DE}, \underline{\neg m^1_{CE}}, C^0, \underline{\neg d^0}, \underline{\neg e^0}) = 0$ (Underlined CPT attributes uniquely identify these entries to be 0). By definition, we obtain $P(\neg e^2|\ldots, e^2, \ldots) = 0$, and, thus, we stand in conflict with the probability axioms of Kolmogorov.

5 Discussion

Using an ADBN has the benefit of anticipating indirect causes in-time in an over-the-time evolving process. Still, it comes with a cost of introduced activators, which need to be defined and enforce minimal observation sets of activators (Eq. 3: any

acyclic constellation of probably active activators is allowed). However, we came from a point of view where activators existed and Prop. 1 shows that classic DBNs are significantly more restricted (the number of uniformly directed bipartite graphs with n vertices is far smaller than the number of possible DAGs). The need to anticipate indirect influences originated from coarse observation timesteps, where indirect influences must be anticipated to explain made observations. Where we motivated coarse timesteps from an unavailability of finer observations, the choice of coarser timesteps is also motivated by computational feasibility. Not being bound to the finest available observation granularity relaxes the rate of needed time-updates and is also discussed by [Pfeffer and Tai, 2005] in the form of Asynchronous DBNs using [Nodelman *et al.*, 2002]'s CTBNs. Still, Asynchronous DBNs and CTBNs run into the same problem given in Prop. 1 of anticipating indirect influences during one timestep.

We have discussed the complexity of operations over *time* and have shown that in an ADBN we obtain the same, and even simpler, complexities as in classic DBNs. However, like in any other DBN, the dimension complexity in terms of nodes of one operation remains computationally intractable and demands approximate inference techniques, which can greatly benefit from context specific independencies, as shown by [Boutilier *et al.*, 1996].

Notwithstanding, it is possible that Eq. 3 does not hold in a particular situation. Still, we now have a direct indicator for potentially spurious results. In this particular situation a mitigation is needed, but is beyond the scope of this paper. Such a mitigation would be, for example, to move small subsets of activators to a neighboring timestep or enforcing observations of more deactive activators.

6 Conclusion

We have shown that indirect causes in dynamic Bayesian networks cause conflicts in representing causality. These conflicts arose from using a modeled dimension for assuring syntactic requirements. By extending dynamic Bayesian networks with activator variables to ADBNs, we are able to move acyclicity constraints from a design phase to a later operation phase. Without the need of algorithm frameworks, degrading a Bayesian network to a reasoning process, we obtained a solid mathematical basis sound to Bayesian networks with a causally correct anticipation of indirect causes in dynamic Bayesian networks under much softer restrictions.

Future work is dedicated to further acyclicity constraints when considering properties of local CPDs, which even allow operations with cyclic activator sets and extensions to relational Bayesian networks [Jaeger, 1997].

References

[Boutilier *et al.*, 1996] Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller. Context-Specific Independence in Bayesian Networks. In *Twelfth Conference on Uncertainty in Artificial Intelligence*, volume 4, pages 115–123, 1996.

[Ferrucci *et al.*, 2010] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. Building Watson: An Overview of the DeepQA Project. *AI magazine*, 31(3):59–79, 2010.

[Glesner and Koller, 1995] Sabine Glesner and Daphne Koller. Constructing Flexible Dynamic Belief Networks from First-Order Probabilistic Knowledge Bases. *Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, 946:217–226, 1995.

[Haddawy *et al.*, 1995] Peter Haddawy, James Helwig, Liem Ngo, and Robert Krieger. Clinical Simulation using Context-Sensitive Temporal Probability Models. In *Symposium on Computer Applications in Medical Care*, volume 1, pages 203–207, 1995.

[Jaeger, 1997] Manfred Jaeger. Relational Bayesian Networks. In *Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 266–273, 1997.

[Jaeger, 2001] Manfred Jaeger. Complex Probabilistic Modeling with Recursive Relational Bayesian Networks. *Annals of Mathematics and Artificial Intelligence*, 32:179–220, 2001.

[Murphy *et al.*, 2014] Kevin Murphy, Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In *20th International Conference on Knowledge Discovery and Data Mining*, pages 601–610. ACM, 2014.

[Murphy, 2002] Kevin Patrick Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2002.

[Ngo and Haddawy, 1997] Liem Ngo and Peter Haddawy. Answering Queries from Context-Sensitive Probabilistic Knowledge Bases. *Theoretical Computer Science*, 171(1-2):147–177, 1997.

[Ngo *et al.*, 1995] Liem Ngo, Peter Haddawy, and James Helwig. A Theoretical Framework for Context-Sensitive Temporal Probability Model Construction with Application to Plan Projection. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 419–426, 1995.

[Nodelman *et al.*, 2002] Uri Nodelman, Christian R Shelton, and Daphne Koller. Continuous Time Bayesian Networks. In *Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 378–387. Morgan Kaufmann, 2002.

[Pearl, 2002] Judea Pearl. Reasoning with Cause and Effect. *AI Magazine*, 23(1):1–83, 2002.

[Pearl, 2009] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[Pfeffer and Tai, 2005] Avi Pfeffer and Terry Tai. Asynchronous Dynamic Bayesian Networks. In *21st Conference on Uncertainty in Artificial Intelligence*, 2005.

[Sanghai *et al.*, 2005] Sumit Sanghai, Pedro Domingos, and Daniel Weld. Relational Dynamic Bayesian Networks. *Journal of Artificial Intelligence Research*, 24:759–797, 2005.