Approximating Bayesian Belief Networks by Arc Removal

*Robert A. van Engelen**
Dept. of Computer Science, Leiden University,
P.O. Box 9512, 2300 RA Leiden, The Netherlands
e-mail: `robert@cs.leidenuniv.nl`

Abstract

Bayesian belief networks or causal probabilistic networks may reach a certain size and complexity where the computations involved in exact probabilistic inference on the network tend to become rather time consuming. Methods for approximating a network by a simpler one allow the computational complexity of probabilistic inference on the network to be reduced at least to some extend. We propose a general framework for approximating Bayesian belief networks based on model simplification by arc removal. The approximation method aims at reducing the computational complexity of probabilistic inference on a network at the cost of introducing a bounded error in the prior and posterior probabilities inferred. We present a practical approximation scheme and give some preliminary results.

1 Introduction

Today, more and more applications based on the Bayesian belief network[1] formalism are emerging for reasoning and decision making in problem domains with inherent uncertainty. Current applications range from medical diagnosis and prognosis [1], computer vision [10], to information retrieval [2]. As applications grow larger, the belief networks involved increase in size. And as the topology of the network becomes more dense, the run-time complexity of probabilistic inference increases dramatically, reaching a state where real-time decision making eventually becomes prohibitive; exact inference in general with Bayesian belief networks has been proven to be NP-hard [3].

For many applications, computing exact probabilities from a belief network is liable to be unrealistic due to inaccuracies in the probabilistic assessments for the network. Therefore, in general, approximate methods suffice. Furthermore, the employment of approximate methods alleviates probabilistic inference on a network at least to some extend. Approximate methods provide probability estimates either by employing simulation methods for approximate inference, first introduced by Henrion [7], or through methods based on model simplification, examples are *annihilating small probabilities* [8] and *removal of weak dependencies* [13].

With the former approach, *stochastic simulation* methods [4] provide for approximate inference based on generating multisets of configurations of all the variables from a belief network. From this multiset, (conditional) probabilities of interest are estimated from the occurrence frequencies. These probability estimates tend to approximate the true probabilities

*Part of this work has been done at Utrecht University, Dept. of Computer Science, The Netherlands.

[1]In this paper we adopt the term *Bayesian belief network* or *belief network* for short. Belief networks are also known as *probabilistic networks*, *causal networks*, and *recursive models*.

if the generated multiset is sufficiently large. Unfortunately, the computational complexity of approximate methods is still known to be NP-hard [5] if a certain accuracy of the probability estimates is demanded for. Hence, just like exact methods, simulation methods have an exponential worst-case computational complexity.

As has been demonstrated by Kjaerulff [13], forcing additional conditional independence assumptions portrayed by a belief network provides a promising direction towards belief network approximation in view of model simplification. However, Kjaerulff's method is specifically tailored to the *Bayesian belief universe* approach to probabilistic inference [9] and model simplification is not applied to a network directly but to the belief universes obtained from a belief network. The method identifies weak dependencies in a belief universe of a network and removes these by removing specific links from the network thereby enforcing additional conditional independencies portrayed by the network. As a result, a speedup in probabilistic inference is obtained at a cost of a bounded error in inference.

In this paper we propose a general framework for belief network approximation by arc removal. The proposed approximation method adopts a similar approach as Kjaerulff's method [13] with respect to the means for quantifying the strength of arcs in a network in terms of the Kullback-Leibler information divergence statistic. In general, the Kullback-Leibler information divergence statistic [14] provides a means for measuring the divergence between a probability distribution and an approximation of the distribution, see e.g. [22]. However, there are important differences to be noted between the approaches. Firstly, the type of independence statements enforced in our approach renders the *direct* dependence relationship portrayed by an arc superfluous, in contrast to Kjaerulff's method where other links may be rendered superfluous as well. As a consequence, we apply more localized the changes to the network which allows a large set of arcs to be removed simultaneously. Secondly, as has been mentioned above, Kjaerulff's method operates only with the Bayesian belief universe approach to probabilistic inference using the clique-tree propagation algorithm of Lauritzen and Spiegelhalter [16]. In contrast, the framework we propose operates on a network directly and therefore applies to any type of method for probabilistic inference. Finally, given an upper bound on the posterior error in probabilistic inference allowed, a (possibly large) set of arcs is removed simultaneously from a belief network requiring only one pre-evaluation of the network in contrast to Kjaerulff's method in which conditional independence assumptions are added to the network one at a time.

The rest of this paper is organized as follows. Section 2 provides some preliminaries from the Bayesian belief network formalism and introduces some notions from information theory. In Section 3, we present a method for removing arcs from a belief network and analyze the consequences of the removals on the represented joint probability distribution. In Section 4, some practical approximation schemes are discussed, aimed at reducing the computational complexity of inference on a belief network. To conclude, in Section 5 the advantages and disadvantages of the presented method are compared to other existing methods for approximating belief networks.

2 Preliminaries

In this section we briefly review the basic concepts of the Bayesian belief network formalism and some notions from information theory. In the sequel, we assume that the reader is well acquainted with probability theory and with the basic notions from graph theory.

2.1 Bayesian Belief Networks

Bayesian belief networks allow for the explicit representation of dependencies as well as independencies using a graphical representation of a joint probability distribution. In general, undirected and directed graphs are powerful means for representing independency models, see e.g. [21, 22]. Associated with belief networks are algorithms for *probabilistic inference* on a network by propagating *evidence*, providing a means for reasoning with the uncertain knowledge represented by the network.

A belief network consists of a qualitative and a quantitative representation of a joint probability distribution. The qualitative part takes the form of an acyclic digraph G in which each vertex $V_i \in V(G)$ represents a discrete statistical variable for stating the truth of a proposition within a problem domain. In the sequel, the notions of vertex and variable are used interchangeably. Each arc in the digraph, which we denote as $V_i \to V_j \in A(G)$ between vertex V_i, called the *tail* of the arc, and vertex V_j, called the *head* of the arc, represents a direct causal influence between the vertices discerned. Then, vertex V_i is called an immediate *predecessor* of vertex V_j and vertex V_j is called an immediate *descendant* of vertex V_i. Furthermore, associated with the digraph of a belief network is a numerical assessment of the strengths of the causal influences, constituting the quantitative part of the network.

In the sequel, for ease of exposition, we assume binary statistical variables taking values in the domain $\{\text{TRUE}, \text{FALSE}\}$. However, the generalization to variables taking values in any finite domain is straightforward. Each variable V_i represents a proposition where $V_i = \text{TRUE}$ is denoted as v_i and $V_i = \text{FALSE}$ is denoted as $\neg v_i$. For a set of variables V, the conjunction $C_V = \bigwedge_{V_i \in V} V_i$ of all variables $V_i \in V$ is called the *configuration scheme* of V; a *configuration* c_V of V is a conjunction of value assignments to the variables in V. In the sequel, we use the concept of configuration scheme to denote that a specific property holds for all possible configurations of a set of variables.

Definition 2.1 *A* Bayesian belief network *is a tuple* $B = (G, \Gamma)$ *where*

- $G = (V(G), A(G))$ *is an acyclic digraph with* $V(G) = \{V_1, \ldots, V_n\}$, $n \geq 1$, *and*
- $\Gamma = \{\gamma_{V_i} \mid V_i \in V(G)\}$ *is a set of real-valued functions* $\gamma_{V_i}\colon \{C_{V_i}\} \times \{C_{\pi_G(V_i)}\} \to [0,1]$, *called* (conditional) probability assessment functions, *such that for each configuration* $c_{\pi_G(V_i)}$ *of the set* $\pi_G(V_i)$ *of immediate predecessors of vertex* V_i *we have that* $\gamma_{V_i}(\neg v_i \mid c_{\pi_G(V_i)}) = 1 - \gamma_{V_i}(v_i \mid c_{\pi_G(V_i)})$, $i = 1, \ldots, n$.

A probabilistic meaning is assigned to the topology of the digraph of a belief network by means of the *d-separation criterion* [18]. The criterion allows for the detection of dependency relationships between the vertices of the network's digraph by traversing undirected paths, called *chains*, comprised by the directed links in the digraph. Chains can be *blocked* by a set of vertices as is stated more formally in the following definition.

Definition 2.2 *Let* $G = (V(G), A(G))$ *be an acyclic digraph. Let* ξ *be a chain in* G. *Then* ξ *is* blocked *by a set of vertices* $W \subseteq V(G)$ *if* ξ *contains three consecutive vertices* X_1, X_2, $X_3 \in W$ *for which one of the following three conditions is fulfilled:*

- $X_1 \leftarrow X_2$ *and* $X_2 \to X_3$ *are on the chain* ξ *and* $X_2 \in W$;
- $X_1 \to X_2$ *and* $X_2 \to X_3$ *are on the chain* ξ *and* $X_2 \in W$;

- *$X_1 \to X_2$ and $X_2 \leftarrow X_3$ are on the chain ξ and $\sigma^*_G(X_2) \cap W = \emptyset$ where $\sigma^*_G(X_2)$ denotes the set of vertices composed of X_2 and all its descendants.*

Note that a chain ξ is blocked by $\emptyset$ if and only if ξ contains $X_1 \to X_2$ and $X_2 \leftarrow X_3$. In this case, vertex X_2 is called a *head-to-head* vertex with respect to ξ [6].

Definition 2.3 *Let $G = (V(G), A(G))$ be an acyclic digraph and let X, Y, $Z \subseteq V(G)$ be disjoint subsets of vertices from G. The set Y is said to* d-separate *the sets X and Z in G, denoted $\langle X \mid Y \mid Z\rangle^d_G$, if for each $V_i \in X$ and $V_j \in Z$ every chain from V_i to V_j in G is blocked by Y.*

The d-separation criterion provides for the detection of probabilistic independence relations from the digraph of a belief network, as is stated more formally in the following definition.

Definition 2.4 *Let $G = (V(G), A(G))$ be an acyclic digraph. Let $\Pr$ be a joint probability distribution on $V(G)$. Digraph G is an* I-map *for $\Pr$ if $\langle X \mid Z \mid Y\rangle^d_G$ implies $X \perp\!\!\!\perp_{\Pr} Y \mid Z$ for all disjoint subsets X, Y, $Z \subseteq V(G)$, i.e. X is conditionally independent of Z given Y in $\Pr$.*

By the chain-rule representation of a joint probability distribution from probability theory, the initial probability assessment functions of a belief network provide all the information necessary for uniquely defining a joint probability distribution on the set of variables discerned that respects the independence relations portrayed by the digraph [11, 18].

Theorem 2.5 *Let $B = (G, \Gamma)$ be a belief network as defined in Definition 2.1. Then,*

$$\Pr(C_{V(G)}) = \prod_{V_i \in V(G)} \gamma_{V_i}(V_i \mid C_{\pi_G(V_i)})$$

defines a joint probability distribution $\Pr$ on $V(G)$ such that G is an I-map for $\Pr$.

A belief network therefore uniquely represents a joint probability distribution. For computing (conditional) probabilities from a network, several efficient algorithms have been developed from which Pearl's *polytree algorithm* with *cutset conditioning* [18, 19] and the method of *clique-tree propagation* by Lauritzen and Spiegelhalter [16] (and combinations [20]) are the most widely used algorithms for *exact* probabilistic inference. Simulation methods provide for *approximate* probabilistic inference, see [4] for an overview.

2.2 Information Theory

The *Kullback-Leibler information divergence* [14] has several important applications in statistics. One of which is for measuring how well one joint probability distribution can be approximated by another with a simpler dependence structure, see e.g. [22]. In the sequel, we will make extensive use of the Kullback-Leibler information divergence. Before defining the Kullback-Leibler information divergence more formally, the concept of continuity is introduced [14].

Definition 2.6 *Let V be a set of statistical variables and let $\Pr$ and $\Pr'$ be joint probability distributions on V. Then $\Pr$ is* absolutely continuous *with respect to $\Pr'$ over a subset of variables $X \subseteq V$, denoted as $\Pr \ll \Pr' \parallel X$, if $\Pr(c_X) = 0$ whenever $\Pr'(c_X) = 0$ for all configurations c_X of X.*

We will write $\Pr \ll \Pr'$ for $\Pr \ll \Pr' \parallel V$ for short. Note that the continuity relation is a reflexive and transitive relation on probability distributions. Furthermore, the continuity relation satisfies

- if $\Pr \ll \Pr' \parallel X$, then $\Pr \ll \Pr' \parallel Y$ for all subsets of variables $X, Y \subseteq V$ with $Y \subseteq X$;
- if $\Pr \ll \Pr' \parallel X$, then $\Pr(\cdot \mid c_Y) \ll \Pr'(\cdot \mid c_Y) \parallel X$ for all subsets of variables $X, Y \subseteq V$ and each configuration c_Y of Y with $\Pr(c_Y) > 0$.

That is, if a joint probability distribution Pr is absolutely continuous with respect to a distribution $\Pr'$ over some set of variables X, then Pr is also absolutely continuous with respect to $\Pr'$ over any subset of X. In addition, any posterior distribution $\Pr(\cdot \mid c_Y)$ of Pr given some configuration c_Y of Y is also absolutely continuous with respect to the posterior distribution $\Pr'(\cdot \mid c_Y)$ of $\Pr'$ given c_Y over X.

Definition 2.7 *Let V be a set of statistical variables and let $X \subseteq V$. Let* Pr *and* $\Pr'$ *be joint probability distributions on V. The* Kullback-Leibler information divergence *or* cross entropy *of* Pr *with respect to* $\Pr'$ *over X, denoted as $I(\Pr, \Pr'; X)$, is defined as*

$$I(\Pr, \Pr'; X) = \begin{cases} \sum\limits_{c_X} \Pr(c_X) \cdot \log \dfrac{\Pr(c_X)}{\Pr'(c_X)} & \text{if } \Pr \ll \Pr' \parallel X \\ \infty & \text{otherwise} \end{cases}$$

where $0 \cdot \log(0/\Pr'(c_X)) = 0$.

In the sequel, we will write $I(\Pr, \Pr')$ for $I(\Pr, \Pr'; V)$ for short. Note that the information divergence is not symmetric in Pr and $\Pr'$ and is finite if and only if Pr is absolutely continuous with respect to $\Pr'$. Furthermore, the information divergence I satisfies

- $I(\Pr, \Pr'; X) \geq 0$ for all subsets of variables $X \subseteq V$, especially $I(\Pr, \Pr'; X) = 0$ if and only if $\Pr(C_X) = \Pr'(C_X)$;
- $I(\Pr, \Pr'; X) \leq I(\Pr, \Pr'; V)$ for all subsets of variables $X \subseteq V$; and
- $I(\Pr, \Pr'; X \cup Y) = I(\Pr, \Pr'; X) + I(\Pr, \Pr'; Y)$ for all subsets of variables $X, Y \subseteq V$ if X and Y are independent in both Pr and $\Pr'$.

In principle, the base of the logarithm for the Kullback-Leibler information divergence is immaterial, providing only a unit of measure; in the sequel, we use the natural logarithm. With this assumption the following property holds.

Proposition 2.8 *Let V be a set of statistical variables and let* Pr *and* $\Pr'$ *be joint probability distributions on V. Furthermore, let I be the Kullback-Leibler information divergence as defined in Definition 2.7. Then,*

$$|\Pr(C_X) - \Pr'(C_X)| \leq \sqrt{\frac{1}{2} I(\Pr, \Pr'; V)}$$

for all $X \subseteq V$.

Hence, the Kullback-Leibler information divergence provides for an upper bound on the *absolute divergence* $|\Pr(c_X) - \Pr'(c_X)|$ over all configurations c_X of X, a property of the Kullback-Leibler information divergence known as the *information inequality* [15].

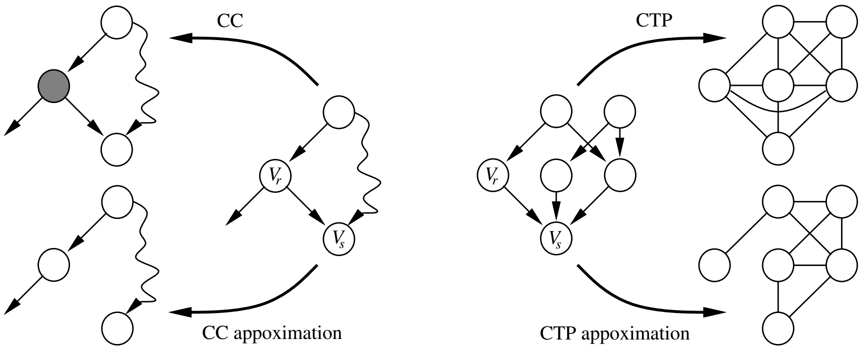

Figure 1: Reducing the complexity of cutset conditioning (CC) and clique-tree propagation (CTP) by removing arc $V_r \to V_s$.

3 Approximating a Belief Network by Removing Arcs

In this section we propose a method for removing arcs from a belief network and we investigate the consequences of the removal on the computational resources and the error introduced. For ease of exposition, a method for removing a single arc from a belief network is introduced first. Then, based on this method and the observations made, a method for multiple simultaneous arc removals is presented.

3.1 Reducing the Complexity of a Belief Network by Removing Arcs

The computational complexity of exact probabilistic inference on a belief network depends to a large extend on the connectivity of the digraph of the network. Removing an arc from the digraph of the network may substantially reduce the complexity of probabilistic inference on the network. For Pearl's polytree algorithm with the method of cutset conditioning [18, 19], undirected cycles, called *loops* [18], can be broken resulting in smaller loop *cutsets* to be used. The size of the cutset determines the computational complexity of inference on the network to a large extend. For the method of clique-tree propagation [16], a belief network is first transformed into a *decomposable graph*. Here, the computational complexity of inference depends to a large extend on the size of the largest clique in the decomposable graph. Removal of an appropriate arc or edge results in splitting cliques into several smaller cliques, see e.g. the method of Kjaerulff [13], yielding a reduction in computational complexity of inference on the decomposable graph.

In Figure 1 we have depicted the effect of removing an arc from the digraph of a belief network for the method of cutset conditioning and for the method of clique-tree propagation. For cutset conditioning, a vertex in the cutset (e.g. the vertex drawn in shading) is required to break the loop. Since removal of arc $V_r \to V_s$ breaks the loop, a smaller cutset may be necessary. For clique-tree propagation, the decomposable graph obtained from the example belief network has three cliques, each with 4 vertices. Removal of arc $V_r \to V_s$ results in a decomposable graph with four smaller cliques, one with 2 and three with 3 vertices.

For approximate methods, the computational complexity of for example *forward simulation* [4] depends to some extend on the distance from a root vertex to a leaf vertex. Therefore, the removal of arcs may also yield a reduction in the complexity of approximate inference. However, it is more difficult to analyze and measure the amount of reduction in complexity in general in comparison to exact methods and in the sequel we will discuss arc removal in view of exact methods for probabilistic inference.

3.2 Removing an Arc from a Belief Network

Although several methods for removing an arc from a belief network can be devised, the method for removal of an arc as defined in the following definition is the most natural choice. This will be made clear when we analyze the effects of the removal.

Definition 3.1 *Let $B = (G, \Gamma)$ be a belief network and let* Pr *be the joint probability distribution defined by B. Let $V_r \to V_s \in A(G)$ be an arc in G. We define the tuple $B_{V_r \not\to V_s} = (G_{V_r \not\to V_s}, \Gamma_{V_r \not\to V_s})$ as*

- $G_{V_r \not\to V_s} = (V(G_{V_r \not\to V_s}), A(G_{V_r \not\to V_s}))$ *is the acyclic digraph with* $V(G_{V_r \not\to V_s}) = V(G)$ *and* $A(G_{V_r \not\to V_s}) = A(G) \setminus \{V_r \to V_s\}$;
- $\Gamma_{V_r \not\to V_s} = \{\gamma'_{V_i} \mid V_i \in V(G)\}$ *is the set of functions* $\gamma'_{V_i} : \{C_{V_i}\} \times \{C_{\pi_{G_{V_r \not\to V_s}}(V_i)}\}$ *with* $\gamma'_{V_i} = \gamma_{V_i} \in \Gamma$ *for all* $V_i \in V(G)$, $V_i \neq V_s$, *and* $\gamma'_{V_s}(V_s \mid C_{\pi_{G_{V_r \not\to V_s}}(V_s)}) = \Pr(V_s \mid C_{\pi_G(V_s) \setminus \{V_r\}})$.

Note that network $B_{V_r \not\to V_s} = (G_{V_r \not\to V_s}, \Gamma_{V_r \not\to V_s})$ resulting after removal of an arc $V_r \to V_s$ from the digraph G of a belief network B, again constitutes a belief network. In this network the assessment functions for the head vertex of the arc are changed only. In the sequel, we will refer to $B_{V_r \not\to V_s}$ as the *approximated* belief network after removal of arc $V_r \to V_s$ and the operation of computing $B_{V_r \not\to V_s}$ will be referred to as *approximating* the network.

Removal of an arc from a belief network may result in a change of the represented joint probability distribution. However, the represented dependency structure of the distribution portrayed by the graphical part of the network may be retained by introducing a *virtual* arc between the two vertices for which a physical arc is removed. A virtual arc may serve for the detection of dependencies and independencies in the *original* probability distribution using the d-separation criterion. A virtual arc, however, is not used in probabilistic inference, still allowing for a faster, approximate computation of prior and posterior probabilities from the simplified network.

3.3 The Error Introduced by Removing an Arc

Removing an arc from a belief network yields a (slightly) simplified network that is faster in inference but exhibits errors in the marginal and conditional probability distributions. In this section we will analyze the errors introduced in the prior and posterior distributions upon belief network approximation by removal of an arc. These effects can be summarized as introducing both a change in the qualitative (ignoring any virtual arcs) as well as a change in the quantitative representation of a joint probability distribution.

The Qualitative Error in Prior and Posterior Distributions

The change in the qualitative belief network representation of the probabilistic dependency structure by removing an arc from a belief network is described by the following lemma.

Lemma 3.2 *Let G be an acyclic digraph and let $V_r \to V_s \in A(G)$ be an arc in G. Let $G_{V_r \not\to V_s} = (V(G_{V_r \not\to V_s}), A(G_{V_r \not\to V_s}))$ be the digraph G with arc $V_r \to V_s$ removed, that is, $V(G_{V_r \not\to V_s}) = V(G)$ and $A(G_{V_r \not\to V_s}) = A(G) \setminus \{V_r \to V_s\}$. Then, we have that $\langle \{V_r\} \mid \pi_{G_{V_r \not\to V_s}}(V_s) \mid \{V_s\} \rangle^d_{G_{V_r \not\to V_s}}$.*

Proof. To prove that $\langle \{V_r\} \mid \pi_{G_{V_r \not\to V_s}}(V_s) \mid \{V_s\} \rangle^d_{G_{V_r \not\to V_s}}$ holds, we show that every chain from vertex V_r to vertex V_s in $G_{V_r \not\to V_s}$ is blocked by the set $\pi_{G_{V_r \not\to V_s}}(V_s)$. For such a chain ξ from V_r to V_s two cases can be distinguished:

- ξ comprises an arc $V_i \to V_s$ for some $V_i \in V(G_{V_r \not\to V_s})$, $V_i \neq V_r$. Since $V_i \in \pi_{G_{V_r \not\to V_s}}(V_s)$, chain ξ is blocked by $\pi_{G_{V_r \not\to V_s}}(V_s)$;
- ξ comprises an arc $V_s \to V_j$ for some $V_j \in V(G_{V_r \not\to V_s})$. Since G and, therefore, $G_{V_r \not\to V_s}$ is acyclic, ξ must contain a head-to-head vertex V_k, i.e. a vertex with two converging arcs on ξ. Since $\sigma^*_{G_{V_r \not\to V_s}}(V_k) \cap \pi_{G_{V_r \not\to V_s}}(V_s) = \emptyset$ chain ξ is blocked by $\pi_{G_{V_r \not\to V_s}}(V_s)$.

□

The property states that after removing arc $V_r \to V_s$ from digraph G of a belief network, the simplified graphical representation now yields that variable V_r is conditionally independent of variable V_s given $\pi_{G_{V_r \not\to V_s}}(V_s)$ being the set of immediate predecessors of V_s in the digraph G with arc $V_r \to V_s$ removed.

The Quantitative Error in the Prior Distribution

The change in the qualitative dependency structure portrayed by the network has its quantitative counterpart as the two are inherently linked together in the belief network formalism. To analyze the error of the approximated prior probability distribution, similar to [13, 22] we use the Kullback-Leibler information divergence for a quantitative comparison in terms of the divergence between the joint probability distribution defined by a belief network and the approximated joint probability distribution obtained after removing an arc from the network.

To facilitate the investigation, we will give an expression for the approximated joint probability distribution in terms of the original distribution. First, we will introduce some additional notions related to arcs in a digraph that are useful for describing the properties that follow. These notions are build on the observation that the set of immediate predecessors $\pi_{G_{V_r \not\to V_s}}(V_s)$ d-separates tail vertex V_r from head vertex V_s in the digraph G with arc $V_r \to V_s$ removed.

Definition 3.3 *Let $G = (V(G), A(G))$ be an acyclic digraph and let $V_r \to V_s \in A(G)$ be an arc in G. We define the* arc block *of $V_r \to V_s$ in G, denoted as $\beta_G(V_r \to V_s)$, as the set of vertices $\beta_G(V_r \to V_s) = \pi_G(V_s) \cup \{V_s\}$. Furthermore, we define the* arc environment *of $V_r \to V_s$ in G, denoted as $\eta_G(V_r \to V_s)$, as the set of vertices $\eta_G(V_r \to V_s) = V(G) \setminus \beta_G(V_r \to V_s)$.*

The joint probability distribution defined by the approximated belief network can be factorized in terms of the joint probability distribution defined by the original network.

Lemma 3.4 *Let $B = (G, \Gamma)$ be a belief network and let* Pr *be the joint probability distribution defined by B. Let $V_r \to V_s \in A(G)$ be an arc in G and let $B_{V_r \not\to V_s} = (G_{V_r \not\to V_s}, \Gamma_{V_r \not\to V_s})$ be the* *approximated belief network after removal of $V_r \to V_s$ as defined in Definition 3.1. Then the joint probability distribution* $\Pr_{V_r \not\to V_s}$ *defined by* $B_{V_r \not\to V_s}$ *satisfies*

$$\begin{aligned}\Pr_{V_r \not\to V_s}(C_{V(G)}) &= \Pr(C_{\eta_G(V_r \to V_s)} \mid C_{\beta_G(V_r \to V_s)}) \cdot \Pr(V_r \mid C_{\pi_G(V_s)\setminus\{V_r\}}) \\ &\quad \cdot \Pr(V_s \mid C_{\pi_G(V_s)\setminus\{V_r\}}) \cdot \Pr(C_{\pi_G(V_s)\setminus\{V_r\}})\end{aligned}$$

where $\beta_G(V_r \to V_s)$ is the arc block and $\eta_G(V_r \to V_s)$ is the arc environment of $V_r \to V_s$ in G as defined in Definition 3.3.

Proof. From Theorem 2.5, the joint probability distribution $\Pr_{V_r \not\to V_s}$ defined by network $B_{V_r \not\to V_s}$ equals

$$\Pr_{V_r \not\to V_s}(C_{V(G)}) = \prod_{V_i \in V(G)} \gamma'_{V_i}(V_i \mid C_{\pi_{G_{V_r \not\to V_s}}(V_i)})$$

where $\gamma'_{V_i} \in \Gamma_{V_r \not\to V_s}$ for all $V_i \in V(G)$. Exploiting Definition 3.1 leads to

$$\begin{aligned}\Pr_{V_r \not\to V_s}(C_{V(G)}) &= \gamma'_{V_s}(V_s \mid C_{\pi_{G_{V_r \not\to V_s}}(V_s)}) \cdot \prod_{V_i \in V(G)\setminus\{V_s\}} \gamma_{V_i}(V_i \mid C_{\pi_G(V_i)}) \\ &= \gamma'_{V_s}(V_s \mid C_{\pi_{G_{V_r \not\to V_s}}(V_s)}) \cdot \frac{\Pr(C_{V(G)})}{\gamma_{V_s}(V_s \mid C_{\pi_G(V_s)})}\end{aligned}$$

Now, since $\gamma_{V_s}(V_s \mid C_{\pi_G(V_s)}) = \Pr(V_s \mid C_{\pi_G(V_s)})$ and $\gamma'_{V_s}(V_s \mid C_{\pi_{G_{V_r \not\to V_s}}(V_s)}) = \Pr(V_s \mid C_{\pi_G(V_s)\setminus\{V_r\}})$, we find

$$\begin{aligned}\Pr_{V_r \not\to V_s}(C_{V(G)}) &= \Pr(V_s \mid C_{\pi_G(V_s)\setminus\{V_r\}}) \cdot \frac{\Pr(C_{V(G)})}{\Pr(V_s \mid C_{\pi_G(V_s)})} \\ &= \Pr(V_s \mid C_{\pi_G(V_s)\setminus\{V_r\}}) \cdot \Pr(C_{\eta_G(V_r \to V_s)} \mid C_{\beta_G(V_r \to V_s)}) \cdot \Pr(C_{\pi_G(V_s)}) \\ &= \Pr(C_{\eta_G(V_r \to V_s)} \mid C_{\beta_G(V_r \to V_s)}) \cdot \Pr(V_r \mid C_{\pi_G(V_s)\setminus\{V_r\}}) \\ &\quad \cdot \Pr(V_s \mid C_{\pi_G(V_s)\setminus\{V_r\}}) \cdot \Pr(C_{\pi_G(V_s)\setminus\{V_r\}})\end{aligned}$$

□

Clearly, this property links the graphical implications of removing an arc from a belief network with the numerical probabilistic consequences of the removal; variable V_r is rendered conditionally independent of variable V_s given $\pi_{G_{V_r \not\to V_s}}(V_s)$ after removal of an arc $V_r \to V_s$.

Now, one of the most important consequences to be investigated is the amount of absolute divergence between the prior probability distribution and the approximated distribution. From the information inequality we have

$$|\Pr(C_X) - \Pr_{V_r \not\to V_s}(C_X)| \leq \sqrt{\frac{1}{2} I(\Pr, \Pr_{V_r \not\to V_s})}$$

for all subsets $X \subseteq V$, where Pr and $\Pr_{V_r \not\to V_s}$ are joint probability distributions on the set of variables V defined by a belief network and the network with arc $V_r \to V_s$ removed respectively. However, we recall that this bound is finite only if Pr is absolutely continuous with respect to $\Pr_{V_r \not\to V_s}$. We prove this property in the following lemma.

Lemma 3.5 *Let $B = (G, \Gamma)$ be a belief network. Let $V_r \to V_s \in A(G)$ be an arc in G and let $B_{V_r \not\to V_s} = (G_{V_r \not\to V_s}, \Gamma_{V_r \not\to V_s})$ be the approximated belief network after removal of $V_r \to V_s$ as defined in Definition 3.1. Then the joint probability distribution* Pr *defined by B is absolutely continuous with respect to the joint probability distribution* $\Pr_{V_r \not\to V_s}$ *defined by* $B_{V_r \not\to V_s}$ *over $V(G)$, i.e.* $\Pr \ll \Pr_{V_r \not\to V_s}$.

Proof. To prove that Pr is absolutely continuous with respect to $\Pr_{V_r \not\to V_s}$ over $V(G)$, we prove that $\Pr(c_{V(G)}) > 0$ implies that $\Pr_{V_r \not\to V_s}(c_{V(G)}) > 0$ for all configurations $c_{V(G)}$ of $V(G)$. First observe that from the chain rule of probability theory we have that

$$\Pr(C_{V(G)}) = \Pr(C_{\eta_G(V_r \to V_s)} \mid C_{\beta_G(V_r \to V_s)}) \cdot \Pr(V_r \wedge V_s \mid C_{\pi_G(V_s) \setminus \{V_r\}}) \cdot \Pr(C_{\pi_G(V_s) \setminus \{V_r\}})$$

where $\beta_G(V_r \to V_s)$ is the arc block and $\eta_G(V_r \to V_s)$ is the arc environment of arc $V_r \to V_s$ in G as defined by Definition 3.3. Now consider a configuration $c_{V(G)}$ of $V(G)$ with $\Pr(c_{V(G)}) > 0$. For this configuration we have that $\Pr(c_{\eta_G(V_r \to V_s)} \mid c_{\beta_G(V_r \to V_s)}) > 0$, $\Pr(c_{V_r} \wedge c_{V_s} \mid c_{\pi_G(V_s) \setminus \{V_r\}}) > 0$, and $\Pr(c_{\pi_G(V_s) \setminus \{V_r\}}) > 0$, where $c_{V(G)} = c_{\eta_G(V_r \to V_s)} \wedge c_{\beta_G(V_r \to V_s)} = c_{\eta_G(V_r \to V_s)} \wedge c_{V_r} \wedge c_{V_s} \wedge c_{\pi_G(V_s) \setminus \{V_r\}}$. Furthermore, $\Pr(c_{V_r} \wedge c_{V_s} \mid c_{\pi_G(V_s) \setminus \{V_r\}}) > 0$ implies that $\Pr(c_{V_r} \mid c_{\pi_G(V_s) \setminus \{V_r\}}) > 0$ and $\Pr(c_{V_s} \mid c_{\pi_G(V_s) \setminus \{V_r\}}) > 0$. These observations lead to

$$\begin{aligned} \Pr_{V_r \not\to V_s}(c_{V(G)}) &= \Pr(c_{\eta_G(V_r \to V_s)} \mid c_{\beta_G(V_r \to V_s)}) \cdot \Pr(c_{V_r} \mid c_{\pi_G(V_s) \setminus \{V_r\}}) \\ &\quad \cdot \Pr(c_{V_s} \mid c_{\pi_G(V_s) \setminus \{V_r\}}) \cdot \Pr(c_{\pi_G(V_s) \setminus \{V_r\}}) \\ &> 0 \end{aligned}$$

Hence, if $\Pr(c_{V(G)}) > 0$, then $\Pr_{V_r \not\to V_s}(c_{V(G)}) > 0$ and we conclude that $\Pr \ll \Pr_{V_r \not\to V_s}$. □

From this property of absolute continuity, the Kullback-Leibler information divergence provides a proper upper bound on the error introduced in the joint probability distribution by removal of an arc from the network. However, the bound can be rather coarse as it can be expected that removing an arc may not always affect the prior probabilities of some specific marginal distributions defined by the network. This observation is formalized by the following lemma which states that the divergence in the prior marginal distributions is always zero for sets of vertices that are not descendants of the head vertex of an arc that is removed. In fact, this property is a direct result from the chain-rule representation of the joint probability distribution by a belief network.

Lemma 3.6 *Let $B = (G, \Gamma)$ be a belief network and let* Pr *be the joint probability distribution defined by B. Let $V_r \to V_s \in A(G)$ be an arc in G and let $B_{V_r \not\to V_s} = (G_{V_r \not\to V_s}, \Gamma_{V_r \not\to V_s})$ be the approximated belief network after removal of $V_r \to V_s$ as defined in Definition 3.1. Then the joint probability distribution $\Pr_{V_r \not\to V_s}$ defined by $B_{V_r \not\to V_s}$ satisfies*

$$\Pr_{V_r \not\to V_s}(C_Y) = \Pr(C_Y)$$

for all $Y \subseteq V(G) \setminus \sigma^*_G(V_s)$, where $\sigma^*_G(V_s)$ denotes the set comprised by V_s and all its descendants.

Proof. First, we will prove that

$$\Pr(C_X) = \prod_{V_k \in X} \gamma_{V_k}(V_k \mid C_{\pi_G(V_k)})$$

where $X = V(G) \setminus \sigma^*_G(V_s)$. By applying Theorem 2.5 and by marginalizing Pr we obtain

$$\begin{aligned} \Pr(c_X) &= \sum_{c_{V(G) \setminus X}} \Pr(c_{V(G)}) \\ &= \sum_{c_{V(G) \setminus X}} \prod_{V_j \in V(G)} \gamma_{V_j}(c_{V_j} \mid c_{\pi_G(V_j)}) \\ &= \sum_{c_{V(G) \setminus X}} \prod_{V_j \in V(G)} \left[\prod_{V_j \in V(G) \setminus X} \gamma_{V_j}(c_{V_j} \mid c_{\pi_G(V_j)}) \cdot \prod_{V_k \in X} \gamma_{V_k}(c_{V_k} \mid c_{\pi_G(V_k)}) \right] \end{aligned}$$

for all configurations c_X of X with the assumption that the configurations that occur within the sum adhere to $c_{V(G)} = c_X \wedge c_{\sigma^*_G(V_i)} = \bigwedge_{V_j \in V(G)} c_{V_j}$. Now since $\pi_G(V_k) \cap \sigma^*_G(V_i) = \emptyset$ for all $V_k \in X$ and $\sum_{c_{V(G)\setminus X}} \prod_{V_j \in V(G)\setminus X} \gamma_{V_j}(c_{V_j} \mid c_{\pi_G(V_j)}) = 1$, we find by rearranging terms

$$\begin{aligned}\Pr(c_X) &= \left[\sum_{c_{V(G)\setminus X}} \prod_{V_j \in V(G)\setminus X} \gamma_{V_j}(c_{V_j} \mid c_{\pi_G(V_j)})\right] \cdot \prod_{V_k \in X} \gamma_{V_k}(c_{V_k} \mid c_{\pi_G(V_k)}) \\ &= \prod_{V_k \in X} \gamma_{V_k}(c_{V_k} \mid c_{\pi_G(V_k)})\end{aligned}$$

for all configurations c_X of X. Hence, we have

$$\Pr(C_X) = \prod_{V_k \in X} \gamma_{V_k}(V_k \mid C_{\pi_G(V_k)})$$

By a similar exposition for network $B_{V_r \not\to V_s}$, we have

$$\Pr_{V_r \not\to V_s}(C_X) = \prod_{V_k \in X} \gamma'_{V_k}(V_k \mid C_{\pi_{G_{V_r \not\to V_s}}(V_k)})$$

where $\gamma'_{V_k} \in \Gamma_{V_r \not\to V_s}$. Now observe that from Definition 3.1 $\gamma'_{V_k} = \gamma_{V_k} \in \Gamma$ for all $V_k \in X$ and we obtain $\Pr_{V_r \not\to V_s}(C_X) = \Pr(C_X)$ and since $Y \subseteq X$, by principle of marginalization we conclude that $\Pr_{V_r \not\to V_s}(C_Y) = \Pr(C_Y)$. □

This property provides the key observation for the applicability of multiple arc removals as will be described in Section 3.4.

The Quantitative Error in Posterior Distributions

Belief networks are generally used for reasoning with uncertainty by processing *evidence*. That is, the probability of some hypothesis is computed from the network given some evidence. In the belief network framework, this amounts to computing the revised probabilities from the posterior probability distribution given the evidence. We will investigate the implications on posterior distributions after removal of an arc. We begin our investigation by exploring some general properties of the Kullback-Leibler information divergence.

Lemma 3.7 *Let V be a set of statistical variables and let X, $Y \subseteq V$ be subsets of V. Let* $\Pr$ *and* $\Pr'$ *be joint probability distributions on V. Then the Kullback-Leibler information divergence I satisfies*

$$I(\Pr, \Pr'; X \cup Y) = I(\Pr, \Pr'; Y) + \sum_{c_Y, \Pr(c_Y)>0} \Pr(c_Y) \cdot I(\Pr(\cdot \mid c_Y), \Pr'(\cdot \mid c_Y); X)$$

Proof. We distinguish two cases: the case that $\Pr \ll \Pr' \parallel X \cup Y$ and the case that $\Pr \not\ll \Pr' \parallel X \cup Y$.

- Assume that $\Pr \ll \Pr' \parallel X \cup Y$. This assumption implies that the information divergence $I(\Pr, \Pr'; X \cup Y)$ is finite. By Definition 2.7 we therefore have that

$$\begin{aligned}&I(\Pr, \Pr'; X \cup Y) \\ &= \sum_{c_{X\cup Y}} \Pr(c_{X\cup Y}) \cdot \log \frac{\Pr(c_{X\cup Y})}{\Pr'(c_{X\cup Y})} \\ &= \sum_{c_Y, \Pr(c_Y)>0} \Pr(c_Y) \cdot \left[\sum_{c_X} \Pr(c_X \mid c_Y) \cdot \log \frac{\Pr(c_X \mid c_Y) \cdot \Pr(c_Y)}{\Pr'(c_X \mid c_Y) \cdot \Pr'(c_Y)}\right]\end{aligned}$$

Here, we used the fact that if for some configuration c'_Y of the set of variables Y the probability distribution $\Pr(\cdot \mid c'_Y)$ is undefined, that is, if $\Pr(c'_Y) = 0$, then for any configuration c'_X of X the probability $\Pr(c'_X \wedge c'_Y) = 0$ and, hence, $\Pr(c'_X \wedge c'_Y) \cdot \log(\Pr(c'_X \wedge c'_Y)/\Pr'(c'_X \wedge c'_Y)) = 0$ by definition. Therefore, we let the first sum in the last equality above range over all configurations c_Y of Y for which $\Pr(c_Y) > 0$. Now by rearranging terms we find

$$\begin{aligned} & I(\Pr, \Pr'; X \cup Y) \\ = & \sum_{c_Y, \Pr(c_Y)>0} \Pr(c_Y) \cdot \left[\log \frac{\Pr(c_Y)}{\Pr'(c_Y)} + \sum_{c_X} \Pr(c_X \mid c_Y) \cdot \log \frac{\Pr(c_X \mid c_Y)}{\Pr'(c_X \mid c_Y)} \right] \\ = & \; I(\Pr, \Pr'; Y) + \sum_{c_Y, \Pr(c_Y)>0} \Pr(c_Y) \cdot I(\Pr(\cdot \mid c_Y), \Pr'(\cdot \mid c_Y); X) \end{aligned}$$

Note that $I(\Pr, \Pr'; Y)$ and $I(\Pr(\cdot \mid c_Y), \Pr'(\cdot \mid c_Y); X)$ are finite.

- Assume that $\Pr \not\ll \Pr' \parallel X \cup Y$. This implies that $I(\Pr, \Pr'; X) = \infty$. We will show that $I(\Pr, \Pr'; Y) + \sum_{c_Y, \Pr(c_Y)>0} \Pr(c_Y) \cdot I(\Pr(\cdot \mid c_Y), \Pr'(\cdot \mid c_Y); X) = \infty$. First, observe that from the assumption there exists a configuration $c'_X \wedge c'_Y$ of $X \cup Y$ such that $\Pr(c'_X \wedge c'_Y) > 0$ and $\Pr'(c'_X \wedge c'_Y) = 0$. Now, two cases are distinguished: the case that $\Pr'(c'_Y) = 0$ and the case that $\Pr'(c'_Y) > 0$.
  - Assume that $\Pr'(c'_Y) = 0$. Since $\Pr(c'_X \wedge c'_Y) > 0$ implies that $\Pr(c'_Y) > 0$, this yields that $\Pr \not\ll \Pr' \parallel Y$. By Definition 2.7 $I(\Pr, \Pr'; Y) = \infty$ and using the fact that the divergence I is non-negative $I(\Pr(\cdot \mid c_Y), \Pr'(\cdot \mid c_Y); X) \neq \infty$ which leads to $I(\Pr, \Pr'; Y) + \sum_{c_Y, \Pr(c_Y)>0} \Pr(c_Y) \cdot I(\Pr(\cdot \mid c_Y), \Pr'(\cdot \mid c_Y); X) = \infty$.
  - Assume that $\Pr'(c'_Y) > 0$. From $\Pr(c'_X \wedge c'_Y) > 0$ while $\Pr'(c'_X \wedge c'_Y) = 0$ we get $\Pr(c'_Y) > 0$, $\Pr(c'_X \mid c'_Y) > 0$, and $\Pr'(c'_X \mid c'_Y) = 0$ for the configurations c'_X and c'_Y. Hence, $\Pr(\cdot \mid c'_Y) \not\ll \Pr'(\cdot \mid c'_Y) \parallel X$ and by Definition 2.7 this implies that $\Pr(c'_Y) \cdot I(\Pr(\cdot \mid c_Y), \Pr'(\cdot \mid c_Y); X) = \infty$. Since $I(\Pr, \Pr'; Y)$ is non-negative, we conclude that $I(\Pr, \Pr'; Y) + \sum_{c_Y, \Pr(c_Y)>0} \Pr(c_Y) \cdot I(\Pr(\cdot \mid c_Y), \Pr'(\cdot \mid c_Y); X) = \infty$.

□

This property of the Kullback-Leibler information divergence leads to the following lemma stating an upper bound on the absolute divergence of the posterior probability distribution defined by a belief network given some evidence and the (approximated) posterior probability distribution defined by another (approximated) network.

Lemma 3.8 *Let V be a set of statistical variables and let $\Pr$ and $\Pr'$ be joint probability distributions on V such that $\Pr \ll \Pr'$. Let I be the Kullback-Leibler information divergence. Then,*

$$\left| \Pr(C_X \mid c_Y) - \Pr'(C_X \mid c_Y) \right| \leq \sqrt{\frac{1}{2} \cdot \frac{I(\Pr, \Pr') - I(\Pr, \Pr'; Y)}{\Pr(c_Y)}}$$

for all subsets of variables X, $Y \subseteq V$ and all configurations c_Y of Y with $\Pr(c_Y) > 0$. Furthermore, this upper bound on the absolute divergence is finite.

Proof. Consider two subsets X, $Y \subseteq V$ and a configuration c_Y of Y with $\Pr(c_Y) > 0$. For this configuration, $\Pr \ll \Pr'$ implies that $\Pr'(c_Y) > 0$ and, hence, the posterior distributions $\Pr(\cdot \mid c_Y)$ and $\Pr'(\cdot \mid c_Y)$ are well-defined. Furthermore, since $\Pr \ll \Pr'$ also implies that $\Pr(\cdot \mid c_Y) \ll \Pr'(\cdot \mid c_Y)$ it follows from Proposition 2.8 that we have the finite upper bound

$$|\Pr(C_X \mid c_Y) - \Pr'(C_X \mid c_Y)| \leq \sqrt{\frac{1}{2} \cdot I(\Pr(\cdot \mid c_Y), \Pr'(\cdot \mid c_Y))}$$

Furthermore, Lemma 3.7 yields that

$$\begin{aligned} I(\Pr, \Pr') &= I(\Pr, \Pr'; V \cup Y) \\ &= I(\Pr, \Pr'; Y) + \sum_{c'_Y, \Pr(c'_Y) > 0} \Pr(c'_Y) \cdot I(\Pr(\cdot \mid c_Y), \Pr'(\cdot \mid c_Y)) \end{aligned}$$

When we consider the divergence $I(\Pr(\cdot \mid c_Y), \Pr'(\cdot \mid c_Y))$ in isolation, we have

$$I(\Pr(\cdot \mid c_Y), \Pr'(\cdot \mid c_Y)) \leq \frac{I(\Pr, \Pr') - I(\Pr, \Pr'; Y)}{\Pr(c_Y)}$$

since for any configuration c'_Y of Y with $\Pr(c'_Y) > 0$ the divergence $I(\Pr(\cdot \mid c'_Y), \Pr'(\cdot \mid c'_Y))$ is finite and non-negative. From these observations we finally find the finite upper bound

$$\begin{aligned} |\Pr(C_X \mid c_Y) - \Pr'(C_X \mid c_Y)| &\leq \sqrt{\frac{1}{2} \cdot I(\Pr(\cdot \mid c_Y), \Pr'(\cdot \mid c_Y))} \\ &\leq \sqrt{\frac{1}{2} \cdot \frac{I(\Pr, \Pr') - I(\Pr, \Pr'; Y)}{\Pr(c_Y)}} \end{aligned}$$

□

Now, from this property of the information divergence, the absolute divergence between the posterior distribution given evidence c_Y for a subset of variables Y of a belief network B and the approximated network $B_{V_r \not\to V_s}$ after removal of an arc $V_r \to V_s$ is bounded by

$$|\Pr(C_X \mid c_Y) - \Pr_{V_r \not\to V_s}(C_X \mid c_Y)| \leq \sqrt{\frac{1}{2} \cdot \frac{I(\Pr, \Pr_{V_r \not\to V_s}) - I(\Pr, \Pr_{V_r \not\to V_s}; Y)}{\Pr(c_Y)}}$$

where $\Pr$ is the joint probability distribution defined by B and $\Pr_{V_r \not\to V_s}$ is the joint probability distribution defined by $B_{V_r \not\to V_s}$. This bound is finite since $\Pr$ is absolutely continuous with respect to $\Pr_{V_r \not\to V_s}$. Furthermore, from this bound we find that in the worst case, i.e. $I(\Pr, \Pr_{V_r \not\to V_s}; Y) = 0$, the error in probabilistic inference on an approximated belief network is inversely proportional to the square root of the probability of the evidence; the more unlikely the evidence, the larger the error may be.

3.4 Multiple Arc Removals

In this section we generalize the method of single arc removal from belief networks to a method of multiple simultaneous arc removals, thereby still guaranteeing a finite upper bound on the error introduced in the prior and posterior distributions.

We recall from Definition 3.1 that removing an arc yields an appropriate change of the assessment functions only for the head vertex of the arc to be removed. Therefore, this operation can be applied in parallel for all arcs not sharing the same head vertex. To formalize this requirement, we introduce the notion of a linear subset of arcs of a digraph.

Definition 3.9 *Let $G = (V(G), A(G))$ be an acyclic digraph with the set of vertices $V(G) = \{V_1, \ldots, V_n\}$, $n \geq 1$, of G indexed in ascending topological order. The relation $\prec_G \subseteq A(G) \times A(G)$ on the set of arcs of G is defined as $V_r \to V_s \prec_G V_{r'} \to V_{s'}$ if and only if $s > s'$ for all pairs of arcs $V_r \to V_s$, $V_{r'} \to V_{s'} \in A(G)$ in G. Furthermore, let $A \subseteq A(G)$ be a subset of arcs in G. Then we say that A is* linear *with respect to G if the order $\prec_G$ is a total order on A, that is, either $V_r \to V_s \prec_G V_{r'} \to V_{s'}$ or $V_{r'} \to V_{s'} \prec_G V_r \to V_s$ for each pair of distinct arcs $V_r \to V_s$, $V_{r'} \to V_{s'} \in A$.*

Note that a linear subset of arcs from a digraph contains no pair of arcs that have a head vertex in common. Now, we formally define the simultaneous removal of a linear set of arcs from a belief network.

Definition 3.10 *Let $B = (G, \Gamma)$ be a belief network. Let $A \subseteq A(G)$ be a linear subset of arcs in G. We define the* multiply approximated *belief network, denoted as $B_A = (G_A, \Gamma_A)$, as the network resulting after the simultaneous removal of all arcs A from B by Definition 3.1. That is, we obtain network $B_A = (G_A, \Gamma_A)$ with*

- *$G_A = (V(G_A), A(G_A))$ the digraph with $V(G_A) = V(G)$ and $A(G_A) = A(G) \setminus A$;*
- *$\Gamma_A = \{\gamma'_{V_i} \mid V_i \in V(G)\}$ the set of functions $\gamma'_{V_i} : \{C_{V_i}\} \times \{C_{\pi_{G_A}(V_i)}\}$ with $\gamma'_{V_i} = \gamma_{V_i} \in \Gamma$ for all $V_i \in V(G)$ with $V_j \to V_i \notin A$ for any $V_j \in V(G)$, and $\gamma_{V_i}(V_i \mid C_{\pi_{G_{V_j \not\to V_i}}(V_i)}) = \Pr(V_i \mid C_{\pi_G(V_i) \setminus \{V_j\}})$ for all $V_i \in V(G)$ with $V_j \to V_i \in A$.*

To analyze the error introduced in the prior as well as in the posterior distribution after removal of a linear set of arcs from a belief network, we once more exploit the information inequality. For obtaining a proper upper bound, the essential requirement is that the joint probability distribution defined by the original network is absolutely continuous with respect to the distribution defined by the multiply approximated network. To prove this, we will exploit the ordering relation on the arcs of a digraph as defined above. This ordering relation induces a total order on the arcs of a linear subset of arcs in a digraph and we show that a consecutive removal of arcs from a belief network in arc linear order yields a multiply approximated network. Then, by transitivity of the continuity relation, this directly implies that the joint probability distribution defined by the original network is absolutely continuous with respect to the distribution defined by the multiply approximated network.

Lemma 3.11 *Let $B = (G, \Gamma)$ be a belief network and let $\Pr$ be the joint probability distribution defined by B. Let $A = \{V_{r_1} \to V_{s_1}, \ldots, V_{r_n} \to V_{s_n}\} \subseteq A(G)$, $n = |A| \geq 1$, be a linear subset of arcs in G ordered with respect to $\prec_G$ as defined in Definition 3.9, i.e. for all pairs of arcs $V_{r_i} \to V_{s_i}$, $V_{r_j} \to V_{s_j} \in A$ with $V_{r_i} \to V_{s_i} \prec_G V_{r_j} \to V_{s_j}$ we have that $i < j$. Now, let $B_A = (G_A, \Gamma_A)$ be the multiply approximated belief network after removal of all arcs A as defined in Definition 3.10. Then,*

$$B_A = \left(\cdots \left(B_{V_{r_1} \not\to V_{s_1}} \right)_{V_{r_2} \not\to V_{s_2}} \cdots \right)_{V_{r_n} \not\to V_{s_n}}$$

where each (approximated) network on the right-hand side is approximated by removal of an arc $V_{r_i} \to V_{s_i}$, $i = 1, \ldots, n$, as defined in Definition 3.1.

Proof. The proof is by induction on $n = |A|$, the cardinality of A.

Base case $n = 1$: by definition $B_{\{V_{r_1} \to V_{s_1}\}} = B_{V_{r_1} \not\to V_{s_1}}$.

For $n > 1$ assume that $B_{A\setminus\{V_{r_n}\to V_{s_n}\}} = (\cdots(B_{V_{r_1}\not\to V_{s_1}})_{V_{r_2}\not\to V_{s_2}}\cdots)_{V_{r_{n-1}}\not\to V_{s_{n-1}}}$ holds as the hypothesis for induction. Now, consider arc $V_{r_n} \to V_{s_n} \in A$. Then, by principle of induction, to prove that $B_A = (\cdots(B_{V_{r_1}\not\to V_{s_1}})_{V_{r_2}\not\to V_{s_2}}\cdots)_{V_{r_n}\not\to V_{s_n}}$, we now have to prove that $B_A = (B_{A\setminus\{V_{r_n}\to V_{s_n}\}})_{V_{r_n}\not\to V_{s_n}}$. Obviously, the digraphs obtained after removal of this arc are identical, i.e we have $G_A = (G_{A\setminus\{V_{r_n}\to V_{s_n}\}})_{V_{r_n}\not\to V_{s_n}}$. This leaves us with a proof for the probability assessment functions. First, observe that the simultaneous removal of all arcs A from network B yields network B_A with probability assessment functions $\gamma'_{V_i} \in \Gamma_A$ for all $V_i \in V(G)$ where we have that $\gamma'_{V_{s_n}}(V_{s_n} \mid C_{\pi_{G_A}(V_{s_n})}) = \Pr(V_{s_n} \mid C_{\pi_G(V_{s_n})\setminus\{V_{r_n}\}})$. Now, observe that the removal of arc $V_{r_n}\to V_{s_n}$ from network $B_{A\setminus\{V_{r_n}\to V_{s_n}\}}$ yields probability assessment functions $\gamma''_{V_i} \in (\Gamma_{A\setminus\{V_{r_n}\to V_{s_n}\}})_{V_{r_n}\not\to V_{s_n}}$ for all $V_i \in V(G)$ for which we find that $\gamma''_{V_i} = \gamma'_{V_i} \in \Gamma_A$ for all $V_i \neq V_{s_n} \in V(G)$ and $\gamma''_{V_{s_n}}(V_{s_n} \mid C_{\pi_{G_A}(V_{s_n})}) = \Pr_{A\setminus\{V_{r_n}\to V_{s_n}\}}(V_{s_n} \mid C_{\pi_{G_A}(V_{s_n})})$. So it remains to prove that $\gamma'_{V_{s_n}} = \gamma''_{V_{s_n}}$, or equivalently, that $\Pr(V_{s_n} \mid C_{\pi_G(V_{s_n})\setminus\{V_{r_n}\}}) = \Pr_{A\setminus\{V_{r_n}\to V_{s_n}\}}(V_{s_n} \mid C_{\pi_{G_A}(V_{s_n})})$. Now, observe that from the ordering relation $\prec_G$ we find that all arcs $A\setminus\{V_{r_n} \to V_{s_n}\}$ that are removed from B are 'below' arc $V_{r_n} \to V_{s_n}$ in the digraph G of B, i.e. by assuming an ascending topological order of the vertices this implies that $s_i > s_n$ for all $V_{r_i} \to V_{s_i} \in A\setminus\{V_{r_n} \to V_{s_n}\}$. Hence, $(\pi_G(V_{s_n})\cup\{V_{s_n}\})\cap\sigma^*_G(V_{s_i}) = \emptyset$ for all $V_{r_i} \to V_{s_i} \in A\setminus\{V_{r_n} \to V_{s_n}\}$ and by the induction hypothesis, we can apply Lemma 3.6 for each arc in $A\setminus\{V_{r_n} \to V_{s_n}\}$ to find that $\Pr(V_{s_n} \wedge C_{\pi_G(V_{s_n})\setminus\{V_{r_n}\}}) = \Pr_{A\setminus\{V_{r_n}\to V_{s_n}\}}(V_{s_n} \wedge C_{\pi_{G_A}(V_{s_n})})$. Furthermore, this yields that $\Pr(V_{s_n} \mid C_{\pi_G(V_{s_n})\setminus\{V_{r_n}\}}) = \Pr_{A\setminus\{V_{r_n}\to V_{s_n}\}}(V_{s_n} \mid C_{\pi_{G_A}(V_{s_n})})$. Hence, $\gamma'_{V_{s_n}} = \gamma''_{V_{s_n}}$ and we conclude that $B_A = (\cdots(B_{V_{r_1}\not\to V_{s_1}})_{V_{r_2}\not\to V_{s_2}}\cdots)_{V_{r_n}\not\to V_{s_n}}$. □

As a result of this property of multiple arc removals, the Kullback-Leibler information divergence of the joint probability distribution defined by a belief network with respect to the distribution defined by the multiply approximated network is finite. Furthermore, arc linearity implies the following additive property of the Kullback-Leibler information divergence.

Lemma 3.12 *Let $B = (G, \Gamma)$ be a belief network and let* $\Pr$ *be the joint probability distribution defined by B. Let $A \subseteq A(G)$ be a linear subset of arcs in G and let $B_A = (G_A, \Gamma_A)$ be the multiply approximated belief network after removal of all arcs A as defined in Definition 3.10. Let* $\Pr_A$ *be the joint probability distribution defined by B_A. Then the Kullback-Leibler information divergence I satisfies*

$$I(\Pr, \Pr_A) = \sum_{V_r\to V_s\in A} I(\Pr, \Pr_{V_r\not\to V_s})$$

Proof. First, we prove that $\Pr \ll \Pr_A$. Assume that the arcs in the linear set A are ordered according to the relation $\prec_G$ as defined in Definition 3.9, i.e. for all pairs of arcs $V_{r_i} \to V_{s_i}$, $V_{r_j} \to V_{s_j} \in A$ with $V_{r_i} \to V_{s_i} \prec_G V_{r_j} \to V_{s_j}$ we have that $i < j$. From Lemma 3.5 we find that $\Pr \ll \Pr_{V_{r_1}\not\to V_{s_1}}$, $\Pr_{V_{r_1}\not\to V_{s_1}} \ll (\Pr_{V_{r_1}\not\to V_{s_1}})_{V_{r_2}\not\to V_{s_2}}$, $\ldots$, $(\cdots(\Pr_{V_{r_1}\not\to V_{s_1}})\cdots)_{V_{r_{n-1}}\not\to V_{s_{n-1}}} \ll (\cdots(\Pr_{V_{r_1}\not\to V_{s_1}})\cdots)_{V_{r_n}\not\to V_{s_n}}$. Since $\ll$ is transitive, we conclude that $\Pr \ll \Pr_A$ by application of Lemma 3.11. Now, with this observation we find

$$\begin{aligned} I(\Pr, \Pr_A) &= \sum_{c_{V(G)}} \Pr(c_{V(G)}) \cdot \log \frac{\Pr(c_{V(G)})}{\Pr_A(c_{V(G)})} \\ &= \sum_{c_{V(G)}} \Pr(c_{V(G)}) \cdot \log \frac{\prod_{V_i\in V(G)} \gamma_{V_i}(c_{V_i} \mid c_{\pi_G(V_i)})}{\prod_{V_i\in V(G)} \gamma'_{V_i}(c_{V_i} \mid c_{\pi_{G_A}(V_i)})} \end{aligned}$$

where $\gamma_{V_i} \in \Gamma$ and $\gamma'_{V_i} \in \Gamma_A$ for all $V_i \in V(G)$. Since A is linear, we have for each arc $V_r \to V_s \in A$ a new probability assessment function $\gamma'_{V_s} \in \Gamma_A$, while $\gamma'_{V_i} = \gamma_{V_i} \in \Gamma$ for each

$V_i \in V(G)$ with $V_j \to V_i \notin A$ for any vertex $V_j \in V(G)$. This leads to

$$
\begin{aligned}
I(\Pr, \Pr_A) &= \sum_{c_{V(G)}} \Pr(c_{V(G)}) \cdot \log \frac{\prod_{V_r \to V_s \in A} \gamma_{V_s}(c_{V_s} \mid c_{\pi_G(V_s)})}{\prod_{V_r \to V_s \in A} \gamma'_{V_s}(c_{V_s} \mid c_{\pi_{G_{V_r \not\to V_s}}(V_s)})} \\
&\qquad \cdot \frac{\prod_{V_i \in V(G), V_j \to V_i \notin A} \gamma_{V_i}(c_{V_i} \mid c_{\pi_G(V_i)})}{\prod_{V_i \in V(G), V_j \to V_i \notin A} \gamma_{V_i}(c_{V_i} \mid c_{\pi_G(V_i)})} \\
&= \sum_{c_{V(G)}} \Pr(c_{V(G)}) \cdot \sum_{V_r \to V_s \in A} \log \frac{\gamma_{V_s}(c_{V_s} \mid c_{\pi_G(V_s)})}{\gamma'_{V_s}(c_{V_s} \mid c_{\pi_{G_{V_r \not\to V_s}}(V_s)})} \\
&\qquad \cdot \frac{\prod_{V_i \in V(G), V_j \to V_i \notin A} \gamma_{V_i}(c_{V_i} \mid c_{\pi_G(V_i)})}{\prod_{V_i \in V(G), V_j \to V_i \notin A} \gamma_{V_i}(c_{V_i} \mid c_{\pi_G(V_i)})} \\
&= \sum_{V_r \to V_s \in A} \sum_{c_{V(G)}} \Pr(c_{V(G)}) \cdot \log \frac{\Pr(c_{V(G)})}{\Pr_{V_r \not\to V_s}(c_{V(G)})} \\
&= \sum_{V_r \to V_s \in A} I(\Pr, \Pr_{V_r \not\to V_s})
\end{aligned}
$$

□

Note that linearity of a set of arcs to be removed is a sufficient condition for the property stated above, yet not a necessary one.

From these observations, we have that the information inequality provides a finite upper bound on the error introduced in the prior and posterior distributions of an approximated belief network after simultaneous removal of a linear set of arcs. This bound is obtained by summing the information divergences between the joint probability distribution defined by the network and the approximated distribution after removal of each arc individually from the set of arcs.

Example 1 Consider the belief network $B = (G, \Gamma)$ where G is the digraph depicted in Figure 2.

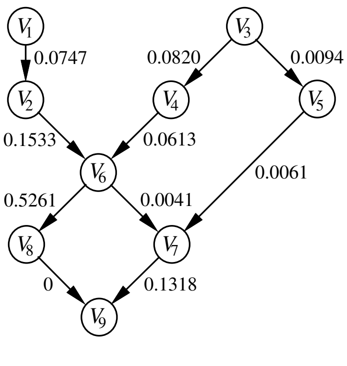

Figure 2: Information divergence for each arc in the digraph of an example belief network.

A	$\sqrt{\frac{1}{2}I(\Pr, \Pr_A)}$	$\max\limits_{c_X, X \subseteq V(G)} \{\lvert \Pr(c_X) - \Pr_A(c_X) \rvert\}$
$\{V_8 \to V_9\}$	0	0
$\{V_6 \to V_7\}$	0.0453	0.0204
$\{V_5 \to V_7\}$	0.0552	0.0190
$\{V_3 \to V_5\}$	0.0686	0.0240
$\{V_6 \to V_7, V_3 \to V_5\}$	0.0822	0.0240
$\{V_5 \to V_7, V_3 \to V_5\}$	0.0880	0.0257
$\{V_4 \to V_6\}$	0.1751	0.0503
$\{V_6 \to V_7, V_4 \to V_6\}$	0.1808	0.0503
$\{V_5 \to V_7, V_4 \to V_6\}$	0.1836	0.0503
$\{V_4 \to V_6, V_3 \to V_5\}$	0.1880	0.0503
$\{V_1 \to V_2\}$	0.1933	0.0840
$\{V_6 \to V_7, V_4 \to V_6, V_3 \to V_5\}$	0.1934	0.0503
$\{V_5 \to V_7, V_4 \to V_6, V_3 \to V_5\}$	0.1960	0.0503
$\{V_6 \to V_7, V_1 \to V_2\}$	0.1985	0.0840

Table 1: Information inequality and absolute divergence of an approximated example belief network.

The set Γ consists of the probability assessment functions $\gamma_{V_1}, \ldots, \gamma_{V_9}$ with

$$
\begin{array}{lll}
\gamma_{V_1}(v_1) = 0.3 & \gamma_{V_6}(v_6 \mid v_2 \wedge v_4) = 0.4 & \gamma_{V_8}(v_8 \mid v_6) = 0.1 \\
\gamma_{V_2}(v_2 \mid v_1) = 0.4 & \gamma_{V_6}(v_6 \mid \neg v_2 \wedge v_4) = 0.6 & \gamma_{V_8}(v_8 \mid \neg v_6) = 1.0 \\
\gamma_{V_2}(v_2 \mid \neg v_1) = 0.8 & \gamma_{V_6}(v_6 \mid v_2 \wedge \neg v_4) = 0.1 & \gamma_{V_9}(v_9 \mid v_7 \wedge v_8) = 0.3 \\
\gamma_{V_3}(v_3) = 0.4 & \gamma_{V_6}(v_6 \mid \neg v_2 \wedge \neg v_4) = 0.9 & \gamma_{V_9}(v_9 \mid \neg v_7 \wedge v_8) = 0.8 \\
\gamma_{V_4}(v_4 \mid v_3) = 0.8 & \gamma_{V_7}(v_7 \mid v_5 \wedge v_6) = 0.6 & \gamma_{V_9}(v_9 \mid v_7 \wedge \neg v_8) = 0.3 \\
\gamma_{V_4}(v_4 \mid \neg v_3) = 0.4 & \gamma_{V_7}(v_7 \mid \neg v_5 \wedge v_6) = 0.4 & \gamma_{V_9}(v_9 \mid \neg v_7 \wedge \neg v_8) = 0.8 \\
\gamma_{V_5}(v_5 \mid v_3) = 0.1 & \gamma_{V_7}(v_7 \mid v_5 \wedge \neg v_6) = 0.6 & \\
\gamma_{V_5}(v_5 \mid \neg v_3) = 0.2 & \gamma_{V_7}(v_7 \mid \neg v_5 \wedge \neg v_6) = 0.5 &
\end{array}
$$

For each arc $V_r \to V_s$ in digraph G, the information divergence $I(\Pr, \Pr_{V_r \not\to V_s})$ between the joint probability distribution Pr defined by B and the joint probability distribution $\Pr_{V_r \not\to V_s}$ defined by the approximated network $B_{V_r \not\to V_s}$ after removal of $V_r \to V_s$ is computed and depicted next to each arc in Figure 2.

Note that despite the presence of arc $V_8 \to V_9$ in G, variables V_8 and V_9 are conditionally independent given variable V_7 from the fact that $\gamma_{V_9}(V_9 \mid V_7 \wedge V_8) = \gamma_{V_9}(V_9 \mid V_7)$. Therefore, this graphically portrayed dependence can be rendered redundant and arc $V_8 \to V_9$ can be removed without introducing an error in the probability distribution since $I(\Pr, \Pr_{V_8 \not\to V_9}) = 0$ as shown in Figure 2.

Table 1 gives the upper bound provided by the information inequality and the absolute divergence of the approximated joint probability distributions after removal of various linear subsets of arcs A from the network's digraph. The table is compressed by leaving out all linear sets containing arc $V_8 \to V_9$ (except for the set $\{V_8 \to V_9\}$) because the second and third column are both unchanged after leaving out this arc. Note that any subset of arcs containing both arcs $V_5 \to V_7$ and $V_6 \to V_7$ is not linear.

From this example, it can be concluded that the upper bound provided by the information inequality exceeds the absolute divergence by a factor of 2 to 3. Furthermore, note that some arcs have more weight in the value of the absolute divergence. For example, the absolute divergence for all sets containing arc $V_4 \to V_6$ is 0.0503. $\diamond$

4 Approximation Schemes

In this section we will present static and dynamic approximation schemes for belief networks. These schemes are based on the observations made in the previous section.

4.1 A Static Approximation Scheme

Clearly, arcs that significantly reduce the computational complexity of inference on a belief network upon removal are most desirable to remove. However, the error introduced upon removal may not be too large. For each arc, the error introduced upon removal of the arc is expressed in terms of the Kullback-Leibler information divergence.

Efficiently Computating the Information Divergence for each Arc

Unfortunately, straightforward computation of the Kullback-Leibler information divergence is computationally far too expensive as it requires summing over all configurations of the entire set of variables, an operation in the order of $\mathcal{O}(2^{|V(G)|})$. However, the following property of the Kullback-Leibler information divergence can be exploited to compute the information divergence locally.

Lemma 4.1 *Let V be a set of statistical variables and let X, Y, $Z \subseteq V$ be mutually disjoint subsets of V. Let $\Pr$ and $\Pr'$ be joint probability distributions on V such that $\Pr'(C_V) = \Pr(C_{V\setminus(X\cup Y\cup Z)} \mid C_{X\cup Y\cup Z}) \cdot \Pr(C_X \mid C_Z) \cdot \Pr(C_Y \mid C_Z) \cdot \Pr(C_Z)$. Then the Kullback-Leibler information divergence I satisfies*

$$I(\Pr, \Pr') = I(\Pr, \Pr'; X \cup Y \cup Z)$$

Proof. By exploiting the factorization of $\Pr'$ in terms of $\Pr$ we find that $\Pr \ll \Pr'$. Using Definition 2.7 we derive

$$\begin{aligned}
& I(\Pr, \Pr') \\
&= \sum_{c_V} \Pr(c_V) \cdot \log \frac{\Pr(c_V)}{\Pr'(c_V)} \\
&= \sum_{c_V} \Pr(c_V) \cdot \log \frac{\Pr(c_{V\setminus(X\cup Y\cup Z)} \mid c_{X\cup Y\cup Z}) \cdot \Pr(c_{X\cup Y\cup Z})}{\Pr(c_{V\setminus(X\cup Y\cup Z)} \mid c_{X\cup Y\cup Z}) \cdot \Pr(c_X \mid c_Z) \cdot \Pr(c_Y \mid c_Z) \cdot \Pr(c_Z)} \\
&= \sum_{c_V} \Pr(c_V) \cdot \log \frac{\Pr(c_{X\cup Y\cup Z})}{\Pr(c_X \mid c_Z) \cdot \Pr(c_Y \mid c_Z) \cdot \Pr(c_Z)} \\
&= \sum_{c_{X\cup Y\cup Z}} \left[\sum_{c_{V\setminus(X\cup Y\cup Z)}} \Pr(c_{V\setminus(X\cup Y\cup Z)} \mid c_{X\cup Y\cup Z}) \right] \cdot \Pr(c_{X\cup Y\cup Z}) \\
&\quad \cdot \log \frac{\Pr(c_{X\cup Y\cup Z})}{\Pr(c_X \mid c_Z) \cdot \Pr(c_Y \mid c_Z) \cdot \Pr(c_Z)}
\end{aligned}$$

Now, since $\sum_{c_{V\setminus(X\cup Y\cup Z)}} \Pr(c_{V\setminus(X\cup Y\cup Z)} \mid c_{X\cup Y\cup Z}) = 1$, this yields that

$$
\begin{aligned}
& I(\Pr, \Pr') \\
&= \sum_{c_{X\cup Y\cup Z}} \Pr(c_{X\cup Y\cup Z}) \cdot \log \frac{\Pr(c_{X\cup Y\cup Z})}{\Pr(c_X \mid c_Z) \cdot \Pr(c_Y \mid c_Z) \cdot \Pr(c_Z)} \\
&= \sum_{c_{X\cup Y\cup Z}} \Pr(c_{X\cup Y\cup Z}) \cdot \log \frac{\Pr(c_{X\cup Y\cup Z})}{\Pr'(c_{X\cup Y\cup Z})} \\
&= I(\Pr, \Pr'; X \cup Y \cup Z)
\end{aligned}
$$

□

For efficiently computing the Kullback-Leibler information divergence $I(\Pr, \Pr_{V_r \not\to V_s})$ for each arc $V_r \to V_s \in A$ of a linear subset of arcs A of the digraph of a belief network, it suffices to sum over all configurations of the arc block $\beta_G(V_r \to V_s)$ of arc $V_r \to V_s$ only, which amounts to computing the quantity

$$
\begin{aligned}
I(\Pr, \Pr_{V_r \not\to V_s}) &= I(\Pr, \Pr_{V_r \not\to V_s}; \beta_G(V_r \to V_s)) \\
&= \sum_{c_{\pi_G(V_s)\cup\{V_s\}}} \gamma_{V_s}(c_{V_s} \mid c_{\pi_G(V_s)}) \cdot \Pr(c_{\pi_G(V_s)}) \\
&\qquad \cdot \log \frac{\gamma_{V_s}(c_{V_s} \mid c_{\pi_G(V_s)}) \cdot \Pr(c_{V_r} \mid c_{\pi_G(V_s)\setminus\{V_r\}})}{\Pr(c_{V_r} \mid c_{\pi_G(V_s)\setminus\{V_r\}}) \cdot \Pr(c_{V_s} \mid c_{\pi_G(V_s)\setminus\{V_r\}})}
\end{aligned}
$$

which is derived by application of the chain rule from probability theory. Hence, the computation of the information divergence $I(\Pr, \Pr_{V_r \not\to V_s})$ only requires the probabilities $\Pr(C_{\pi_G(V_s)})$, $\Pr(V_r \mid C_{\pi_G(V_s)\setminus\{V_r\}})$, and $\Pr(V_s \mid C_{\pi_G(V_s)\setminus\{V_r\}})$ to be computed from the original belief network. In fact, the latter two sets of probabilities can simply be computed from the former set of probabilities using marginalization:

$$
\Pr(V_r \mid C_{\pi_G(V_s)\setminus\{V_r\}}) = \frac{\Pr(C_{\pi_G(V_s)})}{\Pr(v_r \wedge C_{\pi_G(V_s)\setminus\{V_r\}}) + \Pr(\neg v_r \wedge C_{\pi_G(V_s)\setminus\{V_r\}})}
$$

and these conditional probabilities are further used to compute

$$
\begin{aligned}
\Pr(V_s \mid C_{\pi_G(V_s)\setminus\{V_r\}}) &= \gamma_{V_s}(V_s \mid C_{\pi_G(V_s)\setminus\{V_r\}} \wedge v_r) \cdot \Pr(v_r \mid C_{\pi_G(V_s)\setminus\{V_r\}}) \\
&\quad + \gamma_{V_s}(V_s \mid C_{\pi_G(V_s)\setminus\{V_r\}} \wedge \neg v_r) \cdot \Pr(\neg v_r \mid C_{\pi_G(V_s)\setminus\{V_r\}})
\end{aligned}
$$

Furthermore, once the probabilities $\Pr(C_{\pi_G(V_s)})$ are known, the divergence $I(\Pr, \Pr_{V_r \not\to V_s})$ for all arcs $V_r \to V_s$ that share the same head vertex V_s can be computed simultaneously since these computations only require the probabilities $\Pr(C_{\pi_G(V_s)})$.

Selecting a Set of Arcs for Removal

For selecting an optimal set of linear arcs for removal one should carefully weight the advantage of the reduction in computational complexity in inference on a belief network and the disadvantage of the error introduced in the represented joint probability distribution after removal of the arcs.

Given a linear subset of arcs A from the digraph of a belief network B, we define the function expressing the *exact* reduction in computational complexity of inference on network B as

$$
c(B, A) = K(B) - K(B_A)
$$

where K is a cost function expressing the computational complexity of inference on a network. Furthermore, we define the *exact* divergence function d given arcs A on the probability distribution Pr defined by network B as the absolute divergence

$$d(\Pr, A) = \max_{c_X, X \subseteq V(G)} \{|\Pr(c_X) - \Pr_A(c_X)|\}$$

Note that function K depends on the algorithms used for probabilistic inference. For example, if the clique-tree propagation algorithm of Lauritzen and Spiegelhalter is employed, $K(B)$ expresses the sum of the number of configurations of the sets of variables of the cliques of the decomposable graph rendered upon moralization and subsequent triangulation of the digraph. Then, $K(B_A)$ expresses this complexity in terms of the approximated network B after removal of arcs A. Here, we assume an optimal triangulation of the moral graphs of B and B_A, since a bad triangulation of the moral graph of B_A may even yield a negative value for $c(B, A)$. If Pearl's polytree algorithm with cutset conditioning is employed, $K(B)$ equals the number of configurations of the set of variables of the loop cutset of the digraph. Now, an optimal selection method weights the advantage expressed by $c(B, A)$ and disadvantage expressed by $d(\Pr, A)$ for removal of a set of arcs A from network B.

Unfortunately, an optimal selection scheme will first of all depend heavily on the algorithms used for probabilistic inference and, secondly, will depend on the purpose of the network within a specific application. Furthermore, it is rather expensive from a computational point of view to evaluate the exact measures c and d for all possible linear subsets of arcs. In general, the employment of heuristic measures for the selection of a near optimal set of arcs for removal will suffice. To avoid costly evaluations for all possible subsets of arcs, the heuristic measures should be based on combining the *local* advantages (or disadvantages) of removing each arc individually. Such heuristic functions $\tilde{c}$ and $\hat{d}$ for respectively c and d, expressing the impact on the computational complexity and error introduced by removing an arc may be defined with various degrees of sophistication. In fact, the Kullback-Leibler information divergence measures how well one joint probability distribution can be approximated by another exhibiting a simpler dependence structure [22, 13]. Hence, instead of computing the absolute divergence, the information inequality can be used:

$$\hat{d}(\Pr, A) = \sqrt{\frac{1}{2} \sum_{V_r \to V_s \in A} I(\Pr, \Pr_{V_r \not\to V_s}; \beta_G(V_r \to V_s))}$$

where $I(\Pr, \Pr_{V_r \not\to V_s}; \beta_G(V_r \to V_s))$ is the information divergence associated with each arc $V_r \to V_s \in A$ as described in the previous section. Note that $\hat{d}$ now combines the divergence of removing each arc separately and independently.

For defining a heuristic function $\tilde{c}$ valuing the reduction in computational complexity of inference with *exact* methods for probabilistic inference upon removal of a set of arcs from a belief network, the following scheme can be employed. The complexity of methods for exacts inference depends to a large extend on the connectivity of the digraph of a belief network. With each arc $V_i \to V_j \in A(G)$ in the digraph G, a set of loops (undirected cycles), denoted as $loopset(V_i \to V_j)$ is associated. A loopset of an arc consists of all loops in the digraph containing the arc; a loopset of an arc provides local information on the role of the arc in the connectivity of the digraph. This set can be found by a depth-first search for all chains from V_i to V_j in the graph, backtracking for all possibilities and storing the set of vertices found along each chain in the form of bit-vector. Now, we define the heuristic function $\tilde{c}$ as

$$\tilde{c}(B, A) = \left| \bigcup_{V_r \to V_s \in A} \mathit{loopset}(V_r \to V_s) \right| + \alpha \cdot |A|$$

i.e. $\tilde{c}$ expresses the number of *distinct* loops that are broken by removal of a set of arcs from the digraph plus a fraction $\alpha \in (0, 1]$ of the the total number of arcs rendered superfluous. The optimal value for α depends on the algorithm used for exact probabilistic inference.

Now, a combined measure reflecting the trade-off between the advantage $\tilde{c}$ and disadvantage $\hat{d}$ of arc removal may have the form

$$w(B, A) = \lambda\, \tilde{c}(B, A) - \hat{d}(\Pr, A)$$

as suggested by Kjaerulff [13] where λ is chosen such that $\tilde{c}(B, A)$ is comparable to $\hat{d}(\Pr, A)$. Function w expresses the *desirability* of removing a set of arcs from a belief network.

Now suppose that a maximum absolute error $\varepsilon > 0$ is allowed in probabilistic inference on a multiply approximated belief network and further suppose that the probability of the evidence to be processed is never smaller than some constant μ. Observe that from Lemma 3.8 a set of arcs A can be safely removed from the network if $\mu \geq \frac{1}{2} I(\Pr, \Pr_A)/\varepsilon^2$. Hence, an optimal set of arcs can be found for removal if we solve the following optimization problem: maximize $w(B, A)$ for $A \subseteq A(G)$ subject to $\hat{d}(\Pr, A) \leq \varepsilon\sqrt{\mu}$ and A is linear. Note that the constraint $\hat{d}(\Pr, A) \leq \varepsilon\sqrt{\mu}$ ensures that the error in the prior and posterior probability distribution never exceeds ε. This optimization problem can be solved by employing a simulated annealing technique [12], or by using an evolutionary algorithm [17], to find a linear set of arcs for removal that is nearly optimal. A 'real' optimal solution is not appropriate to search for, since only heuristic functions are involved in the search process.

Example 2 Consider once more the belief network from Example 1. Suppose that the probability of evidence to be processed by the approximated belief network does not exceed $\mu = 0.5$ and further suppose that the maximum absolute error allowed for the (conditional) probabilities to be inferred from the approximated network is $\varepsilon = 0.1$.

First, three loops in G can be identified: loop 1 constitutes vertices $\{V_3, V_4, V_5, V_6, V_7\}$, loop 2 constitutes vertices $\{V_6, V_7, V_8, V_9\}$, and loop 3 constitutes vertices $\{V_3, V_4, V_5, V_6, V_7, V_8, V_9\}$. Thus, the loopset of arc $V_6 \to V_7$ is $\{1, 2\}$ and the loopset of arc $V_8 \to V_9$ is $\{2, 3\}$. Now, fix $\lambda = 1$ in w and $\alpha = 1$ in $\tilde{c}$. The following table is obtained for $\hat{d}(\Pr, A) \leq \varepsilon\sqrt{\mu} \approx 0.0707$:

A	$\tilde{c}(B, A)$	$\hat{d}(\Pr, A)$	$w(B, A)$
$\{V_8 \to V_9\}$	3	0	3
$\{V_6 \to V_7\}$	3	0.0453	2.9547
$\{V_8 \to V_9, V_6 \to V_7\}$	5	0.0453	4.9547
$\{V_5 \to V_7\}$	3	0.0552	2.9448
$\{V_8 \to V_9, V_5 \to V_7\}$	5	0.0552	4.9448
$\{V_3 \to V_5\}$	3	0.0686	2.9314
$\{V_8 \to V_9, V_3 \to V_5\}$	5	0.0686	4.9314

The linear set $A = \{V_8 \to V_9, V_6 \to V_7\}$ is the most desirable set of arcs for removal ($w(B, A) = 4.9547$). Note that after removal, the graph G_A is singly connected and, therefore, the network is at least twice as fast for probabilistic inference compared to the original network using either Pearl's polytree algorithm with cutset conditioning or the method of clique-tree propagation.

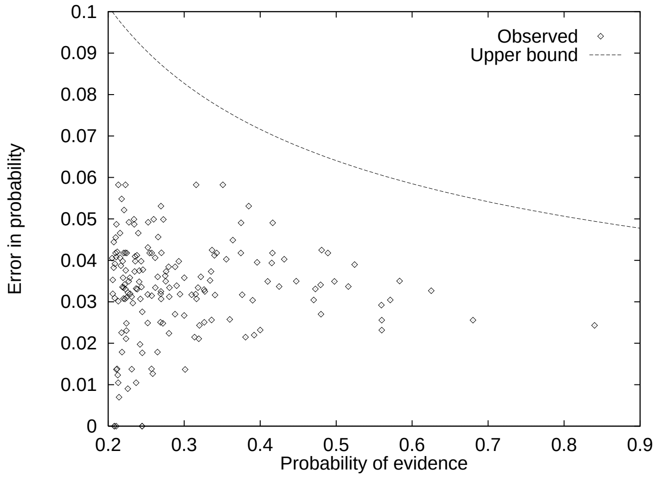

Figure 3: Posterior error in probabilities inferred from an approximated example belief network.

Actually, the probability of evidence that can be processed with the approximated network such that the error in inferred probabilities is bounded by ε requires that $\Pr(c_Y) \geq \frac{1}{2}I(\Pr, \Pr_A)/\varepsilon^2 = 0.205$ for all evidence c_Y, $Y \subseteq V(G)$. In Figure 3 we show the observed maximum absolute error $\max_{c_X, X \subseteq V(G)}\{\Pr(c_X \mid c_Y) - \Pr_A(c_X \mid c_Y)\}$ and upper bound $\sqrt{\frac{1}{2}I(\Pr, \Pr_A)/\Pr(c_Y)}$ obtained for all evidence c_Y, $Y \subseteq V(G)$, with $\Pr(c_Y) \geq 0.205$. $\diamond$

Efficiently Computing an Approximation of a Belief Network

Removal of a linear set of arcs from a belief network requires the computation of new set of probability assessment functions that reflect the introduced qualitative conditional independence with a quantitative conditional independence. We recall from Definition 3.1 that we have that the new probability assessment functions $\gamma'_{V_s}(V_s \mid C_{\pi_{G_{V_r \not\to V_s}}(V_s)}) = \Pr(V_s \mid C_{\pi_G(V_s)\setminus\{V_r\}})$ for variable V_s upon removal of an arc $V_r \to V_s \in A(G)$. Clearly, arc $V_r \to V_s$ is selected for removal only if the Kullback-Leibler information divergence $I(\Pr, \Pr_{V_r \not\to V_s})$ is sufficiently small in order that the error introduced by approximating the network after removal of $V_r \to V_s$ is bounded. The probabilities $\Pr(V_s \mid C_{\pi_G(V_s)\setminus\{V_r\}})$ are in fact already computed by the computation of the information divergence $I(\Pr, \Pr_{V_r \not\to V_s})$ for all arcs $V_r \to V_s$ in the digraph of a belief network. When these probabilities are stored temporarily, it suffices to assign these probabilities to the new probability assessment functions of the head vertex of each arc that is selected for removal.

4.2 A Dynamic Approximation Scheme

In this section we will consider belief networks with singly connected digraphs as a special case for approximation. A singly connected digraph exhibits no loops, that is, at most one chain exists between any two vertices in the digraph. For these networks, arcs can be removed *dynamically* while evidence is being processed in contrast to a *static* removal of arcs as a preprocessing phase before inference as described in the previous section. Therefore, the computational complexity of processing evidence can be reduced depending on the evidence itself and no estimate for a lower bound for the probability of the evidence has to be provided in advance. A detailed description and analysis of the method is beyond the scope of this paper. However, a practical outline of the scheme will be presented which is based on Pearl's polytree algorithm.

First, we will show that all variables in the network retain their prior probabilities upon removal of an arc.

Lemma 4.2 *Let $B = (G, \Gamma)$ be a belief network with a singly connected digraph G. Let $\Pr$ be the joint probability distribution defined by B. Furthermore, let $V_r \to V_s \in A(G)$ be an arc in G and let $B_{V_r \not\to V_s} = (G_{V_r \not\to V_s}, \Gamma_{V_r \not\to V_s})$ be the approximated belief network after removal of $V_r \to V_s$ as defined in Definition 3.1. Let $\Pr_{V_r \not\to V_s}$ be the joint probability distribution defined by $B_{V_r \not\to V_s}$. Then, $\Pr_{V_r \not\to V_s}(V_i) = \Pr(V_i)$ for all variables $V_i \in V(G)$.*

Proof. Assume that the vertices of the singly connected digraph are indexed in ascending topological order, i.e. for each pair of vertices $V_i, V_j \in V(G)$ with a directed path from V_i to V_j in G we have that $i < j$. The proof is by induction on the index i of variable V_i.

Base case $i < s$: from Lemma 3.6 we have that $\Pr_{V_r \not\to V_s}(V_i) = \Pr(V_i)$.

For $i \geq s$, we apply the chain rule and the principle of marginalization to obtain

$$\begin{aligned}
\Pr_{V_r \not\to V_s}(V_i) &= \sum_{c_{\pi_{G_{V_r \not\to V_s}}(V_i)}} \Pr_{V_r \not\to V_s}(V_i \mid c_{\pi_{G_{V_r \not\to V_s}}(V_i)}) \cdot \Pr_{V_r \not\to V_s}(c_{\pi_{G_{V_r \not\to V_s}}(V_i)}) \\
&= \sum_{c_{\pi_{G_{V_r \not\to V_s}}(V_i)}} \gamma'_{V_i}(V_i \mid c_{\pi_{G_{V_r \not\to V_s}}(V_i)}) \cdot \Pr_{V_r \not\to V_s}(c_{\pi_{G_{V_r \not\to V_s}}(V_i)})
\end{aligned}$$

where $\gamma'_{V_i} \in \Gamma_{V_r \not\to V_s}$. Since G is singly connected, all variables $V_j \in \pi_G(V_i)$ are mutually independent by the d-separation criterion. Hence, we have that $\Pr_{V_r \not\to V_s}(c_{\pi_{G_{V_r \not\to V_s}}(V_i)}) = \prod_{V_j \in \pi_{G_{V_r \not\to V_s}}(V_i)} \Pr_{V_r \not\to V_s}(c_{V_j})$. By the assumption that the vertices in G are ordered in ascending topological order, for each $V_j \in \pi_G(V_i)$ we have that $j < i$. Now, by the induction hypothesis assume that for each $V_j \in V(G)$ with $j < i$ we have $\Pr_{V_r \not\to V_s}(V_j) = \Pr(V_j)$. Then, by applying the principle of induction, we find

$$\begin{aligned}
\Pr_{V_r \not\to V_s}(V_i) &= \sum_{c_{\pi_{G_{V_r \not\to V_s}}(V_i)}} \Pr(V_i \mid c_{\pi_{G_{V_r \not\to V_s}}(V_i)}) \cdot \prod_{V_j \in \pi_{G_{V_r \not\to V_s}}(V_i)} \Pr_{V_r \not\to V_s}(c_{V_j}) \\
&= \sum_{c_{\pi_{G_{V_r \not\to V_s}}(V_i)}} \Pr(V_i \mid c_{\pi_{G_{V_r \not\to V_s}}(V_i)}) \cdot \prod_{V_j \in \pi_{G_{V_r \not\to V_s}}(V_i)} \Pr(c_{V_j}) \\
&= \Pr(V_i)
\end{aligned}$$

□

Now, consider an arc $V_r \to V_s$ in a singly connected digraph. In a singly connected digraph no other chain exists from V_r to V_s except for the chain constituting the arc $V_r \to V_s$. Therefore, $\langle \{V_r\} \mid \pi_{G_{V_r \not\to V_s}}(V_s) \cup Y \mid \{V_s\} \rangle^d_{G_{V_r \not\to V_s}}$ holds on the singly connected digraph $G_{V_r \not\to V_s}$ for any subset of variables $Y \subseteq V(G)$. From this observation, we have that the independence relationship between the variables V_r and V_s given $\pi_{G_{V_r \not\to V_s}}(V_s)$ remain unchanged after evidence is given for any subset of variables. Informally speaking, this means that after evidence is processed in a belief network, we can compute the Kullback-Leibler information divergence between the posterior probability distribution defined by a belief network and the posterior distribution of the approximated network after removal of an arc locally. Then, by a similar exposition for the properties of the Kullback-Leibler information divergence applied on general belief networks for multiple arc removals as presented in the previous sections, it can be shown that

$$I(\Pr(\cdot \mid c_Y), \Pr_A(\cdot \mid c_Y)) = \sum_{V_r \to V_s \in A} I(\Pr(\cdot \mid c_Y), \Pr_{V_r \not\to V_s}(\cdot \mid c_Y); \beta_G(V_r \to V_s))$$

for belief network consisting of a singly connected digraph, where Pr is the joint probability distribution defined by the network and $\Pr_A$ is the joint probability distribution defined by the multiple approximated network after removal of all arcs A. We note that the computation of the divergence $I(\Pr(\cdot \mid c_Y), \Pr_{V_r \not\to V_s}(\cdot \mid c_Y); \beta_G(V_r \to V_s))$ for arc $V_r \to V_s$ is as expensive on the computational resources as the computation of the causal and diagnostic messages for vertex V_s in Pearl's polytree algorithm assuming that logarithms require one time unit. Furthermore, in fact, by using Pearl's polytree algorithm, arcs do not have to be physically removed, the blocking of causal and diagnostic messages for updating the probability distribution will suffice. With this observation, we envisage an approximate wave-front version of the polytree algorithm where the sending of messages is blocked between two connected vertices in the graph if the probabilistic dependency relationship between the vertices is very weak. That is, we block all messages for which the information divergence per blocked arc is small such that the total sum of the information divergences over all blocked arcs does not exceed some predetermined constant for the maximum absolute error allowed in probabilistic inference.

5 Discussion and Related Work

We have presented a scheme for approximating Bayesian belief networks based on model simplification through arc removal. In this section we will compare the proposed method with other methods for belief network approximation.

Existing belief network approximation methods are *annihilating small probabilities from belief universes* [8], and *removal of weak dependencies from belief universes* [13]. Both methods have proven to be very successful in reducing the complexity of inference on a belief network on real-life applications using the Bayesian belief universe approach [9].

The method of annihilating small probabilities by Jensen and Andersen reduces the computational effort of probabilistic inference when the method of clique-tree propagation is used for probabilistic inference. The basic idea of the method is to eliminate configurations with small probabilities from belief universes, accepting a small error in the probabilities inferred from the network. To this end, the k smallest probability configurations are selected for each belief universe where k is chosen such that the sum of the probabilities of the selected configurations in the universe is less than some predetermined constant ε. The constant ε determines the maximum error of the approximated prior probabilities. The belief universes are then compressed to take advantage of the zeros introduced. Jensen and Andersen further point out that if the range of probabilities of evidence is known in advance, the method can be applied to approximate a belief network such that the error of the approximated posterior probabilities computed from the network are bounded by some predetermined constant.

Similar to the method of annihilating small probabilities, the method of removal of weak dependencies by Kjaerulff reduces the computational effort of probabilistic inference when the method of clique-tree propagation is used. Kjaerulff's approximation method and the method of annihilation are complementary techniques [13]. The basic idea of the method is to remove edges from the chordal graph constructed from the digraph of a belief network that model weak dependencies. The weaker the dependencies, the smaller the error introduced in the represented joint probability distribution approximated upon removal of an edge. The method operates on the junction tree of a belief network only. Given a constant ε, a set of edges can be removed sequentially such that the error introduced in the prior distribution is smaller than ε. Removal of an edge results in the decomposition of the clique containing the edge into two or more smaller cliques which results in a simplification of the junction tree thereby reducing the computational complexity of inference on the network.

In comparing the methods for approximating belief networks, we first of all find that the method of annihilating small probabilities from belief universes introduces an error that is inversely proportional to the probability of the evidence [8] while the methods based on removing arcs introduces an error that is inversely proportional to the square root of the probability of the evidence. Furthermore, since the original joint probability distribution is absolutely continuous with respect to the approximated probability distribution, the processing of evidence in an approximated belief network by our method is safe in the sense that no undefined conditional probabilities will arise for evidence with a nonzero probability in the original distribution; the evidence that can be processed in an approximated belief network is a superset of the evidence that can be processed in the original network. Once more, this is in contrast to the method of annihilating small probabilities from belief universes. On the other hand, however, the advantage of annihilating small probabilities is that the method operates on the quantitative part of a belief network only whereas arc removal methods change the qualitative representation as well. This can be remedied by introducing virtual arcs to replace removed arcs. Virtual arcs are not used in probabilistic inference.

The method presented in this paper has some similarities to Kjaerulff's method of removal of weak dependencies from belief universes [13]. Both methods aim at reducing inference on a belief network by removing arcs or edges. However, the independency statements we enforce are of the form $V_r \perp\!\!\!\perp V_s \mid \pi_G(V_s) \setminus \{V_r\}$ in contrast to $V_r \perp\!\!\!\perp V_s \mid C \setminus \{V_r, V_s\}$ by Kjaerulff's method where $C \subseteq V(G)$ denotes the clique containing the edge removed by Kjaerulff's method. Furthermore, Kjaerulff's method of removal is based on the clique-tree propagation algorithm only and restricts the removal to one edge from a clique at a time in order that the error introduced is bounded by some predetermined constant. In contrast, our method allows a larger set of arcs (edges) to be removed in parallel, still guaranteeing that the introduced error to be bounded by some predetermined constant regardless of the algorithms for probabilistic inference used.

To summarize the conclusions, the scheme we propose for approximating belief networks operates directly on the digraph of a belief network, has a relatively low computational complexity, provides a bound on the posterior error in the presence of evidence, and is independent of the algorithms used for probabilistic inference.

Acknowledgements

The author would like to acknowledge valuable discussions with Linda van der Gaag of Utrecht University, The Netherlands.

References

[1] S. Andreassen, M. Woldbye, B. Falck, S.K. Andersen, *MUNIN - A Causal Probabilistic Network for Interpretation of Electromyographic Findings*, Proceedings of the Tenth International Joint Conference on Artificial Intelligence, pp. 366–372, 1987.

[2] P.D. Bruza and L.C. van der Gaag, *Index Expression Belief Networks for Information Disclosure*, Journal of Expert Systems: Research and Applications, **7**, pp. 107–138, 1994.

[3] G.F. Cooper, *The Computational Complexity of Probabilistic Inference using Bayesian Belief Networks*, Artificial Intelligence **42**, pp. 393–405, 1990.

[4] S.B. Cousins, W. Chen & M.E. Frisse, *A Tutorial to Stochastic Simulation Algorithms for Belief Networks*, Artificial Intelligence in Medicine **5**, pp. 315–340, 1993.

[5] P. Dagum and M. Luby, *Approximating Probabilistic Inference in Bayesian Belief Networks is NP-hard*, Artificial Intelligence **60**, pp. 141–153, 1993.

[6] D. Geiger, Th. Verma and J. Pearl, *d-Separation: From Theorems to Algorithms*, Technical Report R-130, Computer Science Department, University of California, Los Angeles, 1989.

[7] M. Henrion, *Propagating uncertainty in Bayesian networks by probabilistic logic sampling*, in proceedings Uncertainty in Artificial Intelligence, **4**, North Holland, Amsterdam, pp. 149–163, 1988.

[8] F.V. Jensen and S.K. Andersen, *Approximations in Bayesian Belief Universes for Knowledge-based Systems*, Proceedings of the Sixth Workshop on Uncertainty in Artificial Intelligence, Cambridge, MA, 1990.

[9] F.V. Jensen, S.L. Lauritzen, and K.G. Olesen. *Bayesian updating in causal probabilistic networks by local computations*, Computational Statistics Quarterly, **4**, 1990, pp. 269–282.

[10] F.V. Jensen, J. Nielsen, and H.I. Christensen, *Use of Causal Probabilistic Networks as High LEvel Models in Computer Vision*, Technical Report R-90-39, University of Aalborg, Denmark, 1990.

[11] H. Kiiveri, T.P. Speed, and J.B. Carlin, *Recursive Causal models*, Journal of the Australian Mathematical Society A, pp. 30–52, 1984.

[12] S. Kirkpatrick, C.D. Gelatt, Jr., and M.P. Vecchi, *Optimization by Simulated Annealing*, Science, 220, pp. 671–680, 1983.

[13] U. Kjaerulff, *Reduction of Computational Complexity in Bayesian Networks through Removal of Weak Dependencies*, Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, pp. 374–382, 1994.

[14] S. Kullback, *Information Theory and Statistics*, John Wiley, New York, 1959.

[15] S. Kullback. *A lower bound for discriminating information in terms of variation*, IEEE Transactions on Information Theory **IT-13**, pp. 126–127, 1967.

[16] S.L. Lauritzen and D.J. Spiegelhalter, *Local computations with probabilities on graphical structures and their application to expert systems*, Journal of the Royal Statistical Society, Series B **50**, pp. 157–224, 1988.

[17] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, 2nd ed., Springer-Verlag, 1994.

[18] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo (CA), 1988.

[19] H.J. Suermondt and G.F. Cooper, *Probabilistic inference in multiply connected belief networks using loop cutsets*, Journal of Approximate Reasoning **4**, pp. 283–306, 1990.

[20] H.J. Suermondt and G.F. Cooper, *A Combination of Exact Algorithms for Inference on Bayesian Belief Networks*, Journal of Approximate Reasoning **5**, pp. 521–542, 1991.

[21] N. Wermuth and S.L. Lauritzen, *On Substantive Research Hypothesis, Conditional Independence Graphs and Graphical Chain Models*, Journal of the Royal Statistical Society B, pp. 21–50, 1990.

[22] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons, Inc., Chichester, 1990.