AALBORG UNIVERSITY

Approximation of Bayesian networks through edge removals

by

Uffe Kjærulff

IR-93-2007 ISSN 0908-1216 August 1993

Institute for Electronic Systems
Department of Mathematics and Computer Science
Fredrik Bajers Vej 7E — DK 9220 Aalborg Ø — Denmark
Tel.: +45 98 15 85 22 — TELEX 69 790 aub dk

Approximation of Bayesian networks through edge removals

Uffe Kjærulff

25 August 1993

Abstract

Due to the general inherent complexity of inference in Bayesian networks, the need to compromise the exactitude of inference arises frequently. A scheme for reduction of complexity by enforcing additional conditional independences is investigated. The enforcement of independences is achieved through edge removals in a triangulated graph. The removal of a single edge may imply an enormous reduction of complexity, since other edges may become superfluous by its removal. The approximation scheme presented has several appealing features. Most notably among these, a bound on the overall approximation error can be computed locally, the bound on the error by a series of approximations equals the sum of the bounds of the errors of the individual approximations, and the influence of an approximation attenuates with increasing 'distance' from edge removed. The scheme compares in some cases very favorably with the approximation method suggested by Jensen & Andersen (1990).

Key words. belief network, causal probabilistic network, block recursive graphical model, chain graph, junction tree, hypergraph, discrete Markov field, conditional independence, approximate reasoning, marginal independence graph.

1 Introduction

For domains of knowledge with inherent uncertainty there has been a long tradition among statisticians to model such domains in terms of random Markov fields, where the variables are represented as vertices in a graph and the statistical dependences among the variables are represented as edges of the graph (see e.g. Darroch, Lauritzen & Speed (1980)). Joint probability functions are then specified for clusters of variables. When the dependences have a causal interpretation the edges become directed and a conditional probability function is specified for each variable. In the statistical community, graphical models of this kind is often referred to as recursive graphical models (Wermuth & Lauritzen 1983). With the invention of efficient inference methods for recursive graphical models, first in the singly-connected case (Kim & Pearl 1983) and later in the general case (Lauritzen & Spiegelhalter 1988, Jensen, Lauritzen & Olesen 1990), this knowledge representation scheme has become increasingly popular in the expert systems community. Here, however, the scheme is often termed differently. The most well-known of these are influence diagram (Howard & Matheson 1984), probabilistic causal network (Cooper 1984), Bayesian belief network (Pearl 1986), causal probabilistic network (Andreassen, Woldbye, Falck & Andersen 1987), causal network (Lauritzen & Spiegelhalter 1988). We shall use the term Bayesian network in this paper.

The directed graph of a Bayesian network is acyclic, i.e., it has no directed cycles, and captures all conditional independences between pairs of sets, say A and B, of variables given the values of a third set, C. By exploiting the conditional independence relations among the variables of a network, the underlying joint probability space may be decomposed into smaller subspaces (belief universes) corresponding to the cliques of a triangulated graph (Spiegelhalter 1986, Lauritzen & Spiegelhalter 1988). These belief universes can then be connected in a tree structure called a junction tree (Jensen 1988, Jensen, Lauritzen & Olesen 1990). However, despite the decomposition and despite the success of Bayesian networks in recent years, probabilistic inference in these networks has the inherent disadvantage of being NP-hard (Cooper 1990) and may

hence require excessive computational resources. Thus the need to compromise the exactitude of representation and inference arises frequently. In these cases the complexity of the probabilistic description inhibits or severely restricts representation and inference. Therefore, either more parsimonious methods (if any) for representation and inference should be employed or (further) approximations of the description could take place. We shall concentrate on the second alternative and specifically on approximation under edge removals (i.e., altering the conditional independence properties of a model).

Another scheme for approximating the junction tree representation of Markov models has been described by Jensen & Andersen (1990) who suggested a method based on annihilation of small probabilities by setting the k smallest probabilities to zero for each belief universe of a junction tree, where k is chosen such that the sum of the k smallest probabilities is less than a predetermined value ϵ. After the annihilation, the belief tables of the belief universes are compressed in order to take advantage of the introduced zeros.

Removing an edge has positive and negative consequences. The complexity is reduced on the positive side, whereas, on the negative side, a possibility for erroneous inference is introduced. These two conflicting interests should be reconciled through a trade-off between utility in terms of reduced complexity and loss in terms of possible erroneous inference. To maximize the reduction of complexity we obviously want to remove edges belonging to large belief universes, whereas to minimize the possible error, the dependences neglected (or the edges removed) should be the ones representing the weakest mutual information. As the measure of mutual information we employ the Kullback-Leibler divergence statistic.

It seems obvious that the impact on the system belief by approximating the probability distribution of a belief universe (clique) C cannot be entirely local in general (i.e., restricted to C), but propagates to C's neighbours via the corresponding clique separators. However, if no evidence is received and retransmitted by C after the approximation has been carried out we show that the effect of the approximation does not spread at all, and that for a certain class of edges the effect does not spread even in presence of posterior evidence. For the complementary class of edges the effect propagates throughout the network upon receiving and retransmitting evidence. Intuitively, however, it seems reasonable that the effect of an approximation in C decreases monotonically with increasing distance from C. Proofs are given that evidence received by a clique attenuates monotonically with increasing distance from the clique, and that the Kullback-Leibler divergence is additive over a series of approximations.

The rest of the paper is organized as follows. In Section 2 we briefly review some terminology concerning graphs, Bayesian networks, Markov fields, and junction trees. In Section 3 we characterize the edges which can be removed, and a discussion of the consequences of edge removal in terms of altering the potential representation, the junction tree, and the independence graph is presented in Section 4. In Section 5 we define a criterion for selecting the best edge to remove by reconciling reduction in complexity and possible erroneous inference. We show in Section 6 that a bound on the overall approximation error can be computed locally, that the bound on the error by a series of approximations equals the sum of the bounds of the errors of the individual approximations, and that the influence of an approximation attenuates with increasing 'distance' from edge removed. In Section 7 we compare the edge-removal approach with the approximation method suggested by Jensen & Andersen (1990) by applying the methods on a couple of semi-realistic models. Finally, we conclude the work in Section 8 and discuss possible new areas of research within the presented framework.

2 Terminology

2.1 Graphs and hypergraphs

A (finite) graph $\mathcal{G} = (V, E)$ consists of a finite set of *vertices* V and a set of *edges* $E \subseteq V \times V$ of ordered pairs of distinct vertices. If $(\alpha, \beta) \in E$ and $(\beta, \alpha) \notin E$ the edge from α to β is *directed*, which we shall denote by $\alpha \to \beta$ or $\beta \leftarrow \alpha$. If both $(\alpha, \beta) \in E$ and $(\beta, \alpha) \in E$ the edge between α and β is *undirected* in which case we shall use the shorthand notation $\{\alpha, \beta\} \in E$ or $\alpha — \beta$. A graph is said to be directed if at least one edge

is directed, and undirected if all edges are undirected.

A (finite) *hypergraph* $\mathcal{H} = (V, \mathcal{C})$ consists of a finite set of vertices V and a set of edges $\mathcal{C}$, where $C \subseteq V$ and $|C| \geq 2$ for each $C \in \mathcal{C}$. A hypergraph is an undirected graph if and only if each edge has cardinality two. $\mathcal{C}$ is called a *hypergraph cover* of an undirected graph $\mathcal{G} = (V, E)$ if there is a $C \in \mathcal{C}$ for each $e \in E$ such that $e \subseteq C$. The (undirected) graph $G(\mathcal{C})$ of a hypergraph has vertex set $\bigcup_{C \in \mathcal{C}} C$ and edge set $\{\{\alpha, \beta\} \mid \exists C \in \mathcal{C} : \alpha, \beta \in C\}$.

If there is an edge (directed or undirected) between α and β in a graph $\mathcal{G} = (V, E)$, denoted by $\alpha \sim \beta$, then α and β are *adjacent*. The set of vertices adjacent to α is denoted $\mathrm{adj}(\alpha)$, and the set of edges between α and $\mathrm{adj}(\alpha)$ are called the edges *incident* to α. The set $\mathrm{cl}(\alpha) = \mathrm{adj}(\alpha) \cup \{\alpha\}$ is called the *closure* of α. If $\alpha \to \beta$, then α is called a *parent* of β and β a *child* of α. The sets of parents and children of α are denoted $\mathrm{pa}(\alpha)$ and $\mathrm{ch}(\alpha)$, respectively. If $\mathrm{pa}(\alpha) = \emptyset$, then α is a *founder*. For any subset $A \subseteq V$ we define $\mathrm{pa}(A) = \bigcup_{\alpha \in A} \mathrm{pa}(\alpha) \setminus A$, $\mathrm{ch}(A) = \bigcup_{\alpha \in A} \mathrm{ch}(\alpha) \setminus A$, etc. The subset

$$\mathrm{nb}(\alpha) = \mathrm{adj}(\alpha) \setminus (\mathrm{pa}(\alpha) \cup \mathrm{ch}(\alpha))$$

of $\mathrm{adj}(\alpha)$ is called the *neighbours* of α. The *block* of α is the union of $\{\alpha\}$, $\mathrm{nb}(\alpha)$, $\mathrm{nb}(\beta)$ for all $\beta \in \mathrm{nb}(\alpha)$, $\mathrm{nb}(\gamma)$ for all $\gamma \in \mathrm{nb}(\beta)$, etc, and is denoted $\mathrm{blk}(\alpha)$. In other words, $\mathrm{blk}(\alpha)$ is the set of vertices reachable from α by traversing undirected edges only.

A *path* $\pi = \langle \alpha = \alpha_1, \ldots, \alpha_k = \beta \rangle$ from α to β of length $k - 1$, denoted by $\alpha \rightsquigarrow \beta$, is an ordered sequence of distinct vertices such that $\alpha_i \sim \alpha_{i+1}$ for all $i = 1, \ldots, k-1$; π is *weakly directed*, denoted by $\alpha \rightsquigarrow \beta$, if $(\alpha_i, \alpha_{i+1}) \in E$ for all $i = 1, \ldots, k-1$; π is *(strongly) directed* if it is weakly directed and $\alpha_i \to \alpha_{i+1}$ for some i $(1 \leq i \leq k-1)$; finally, π is *undirected* if it is weakly directed but not strongly directed (i.e., if $\alpha_i — \alpha_{i+1}$ for all $i = 1, \ldots, k-1$). For $A, B, C \subseteq V$, C is said to *separate* A *from* B if for all paths $\langle \alpha = \alpha_1, \ldots, \alpha_k = \beta \rangle$, where $\alpha \in A$ and $\beta \in B$, $\{\alpha_1, \ldots, \alpha_k\} \cap C \neq \emptyset$. A graph $\mathcal{G}$ is *connected* if there is a path between each pair of vertices of $\mathcal{G}$. Unless otherwise stated, connectivity shall be assumed throughout this paper.

An m-cycle is a path $\langle \alpha = \alpha_1, \ldots, \alpha_k = \beta \rangle$ of length m with the exception that $\alpha = \beta$. A *chord* in an m-cycle is an edge $\alpha_i — \alpha_j$ such that $1 < |i - j| < m - 1$ (i.e., $\alpha_i — \alpha_j$ connects the two non-consecutive vertices α_i and α_j of the cycle). A graph $\mathcal{G}$ with no (strongly) directed cycles is called a *chain graph*. Using traditional terminology, $\mathcal{G}$ is called a *directed, acyclic graph* or DAG for short, when all edges are directed. Although a chain graph may also be directed and acyclic using our terminology, we shall stick to the conventional interpretation of the term 'DAG'. Note that for a chain graph $\mathcal{G} = (V, E)$ there exists an ordered k-tuple $(A_1, \ldots, A_k)$ of subsets of V such that (i) for each pair $\alpha, \beta \in A_i$ either $\alpha \not\sim \beta$ or $\alpha — \beta$ for all $i = 1, \ldots, k$, and (ii) for each pair α, β, where $\alpha \in A_i$ and $\beta \in A_j$, either $\alpha \not\sim \beta$ or $\alpha \rightarrow \beta$ for all $1 \leq i < j \leq k$. If there is no $l < k$ such that there exists an ordered l-tuple of subsets of V with that property, then $A_1, \ldots, A_k$ are called *chain components* of $\mathcal{G}$ and can be identified as the connected components generated by the removal of all directed edges from $\mathcal{G}$. If $\mathcal{G}$ is a DAG, then $k = |V|$ (i.e., one vertex per chain component).

In a chain graph, $\mathcal{G} = (V, E)$, the set $\mathrm{de}(\alpha)$ for which $\alpha \rightsquigarrow \beta$ for all $\beta \in \mathrm{de}(\alpha)$ is called the *descendants* of α. The set $\mathrm{an}(\alpha)$ for which $\beta \rightsquigarrow \alpha$ in $\mathcal{G}$ for all $\beta \in \mathrm{an}(\alpha)$ is called the *ancestors* of α. A subset $A \subseteq V$ is an *ancestral set* if $\mathrm{an}(\alpha) \subseteq A$ for all $\alpha \in A$. By $\mathrm{An}(A)$ we denote the smallest ancestral set containing A.

A subset $A \subseteq V$ induces a *subgraph* $\mathcal{G}_A = (A, E_A)$ of $\mathcal{G}$, where $E_A = E \cap (A \times A)$. Note that any subgraph of a chain graph is a chain graph. A graph is *complete* if all vertices are pairwise adjacent. A subset $A \subseteq V$ is complete if it induces a complete subgraph, and if A is maximal (i.e., there is no complete subset $B \subseteq V$ such that $A \subset B$), then it is called a *clique*. A vertex $\alpha \in V$ is *simplicial* if $\mathrm{adj}(\alpha)$ is complete. A hypergraph $\mathcal{H} = (V, \mathcal{C})$ is *conformal* if every clique of $G(\mathcal{C})$ is contained in an edge of $\mathcal{H}$. Thus for any conformal hypergraph with edge set $\mathcal{C}$, $\mathcal{C}$ is the clique set of $G(\mathcal{C})$.

For a chain graph $\mathcal{G} = (V, E)$ with $\mathcal{A}$ being the set of chain components of $\mathcal{G}$, $\mathcal{G}^m$ denotes its *moral graph* obtained in a two-step process: (i) add $\{\alpha, \beta\}$ to E for each pair of distinct vertices $\alpha, \beta \in V$ for which $\alpha \not\sim \beta$ and there exists a chain component $A \in \mathcal{A}$ and a pair of (not necessarily distinct) vertices $\gamma, \delta \in A$ such that $\alpha \rightarrow \gamma$ and $\beta \rightarrow \delta$, and

(ii) make the resulting graph undirected (i.e., for each $(\alpha, \beta) \in E$, where $(\beta, \alpha) \notin E$, add (β, α) to E). For clarity, a moral edge between α and β shall often be denoted $\alpha \overset{\mathrm{m}}{—} \beta$.

An undirected graph $\mathcal{G} = (V, E)$ is *triangulated* (also: *decomposable* or *chordal*) if each cycle of length greater than 3 has a chord. Another way of characterizing triangulated graphs is in terms of sequences of *(proper) decompositions*. A proper decomposition is defined as a triple (A, B, C) of disjoint and non-empty subsets of V such that $V = A \cup B \cup C$, C separates A from B, and C is a complete subset of V. A decomposition (A, B, C) *decomposes* $\mathcal{G}$ into subgraphs $\mathcal{G}_{A \cup C}$ and $\mathcal{G}_{B \cup C}$. A recursive definition of decomposability may now be stated as follows. A graph $\mathcal{G}$ is decomposable either if it is complete or if there exists a decomposition (A, B, C) into decomposable subgraphs $\mathcal{G}_{A \cup C}$ and $\mathcal{G}_{B \cup C}$. Note that the class of triangulated graphs is equivalent to the class of decomposable graphs. For a proof see e.g. Lauritzen, Speed & Vijayan (1984).

Any set T of undirected edges added to an undirected graph $\mathcal{G} = (V, E)$ such that $\mathcal{G}^t = (V, E \cup T)$ is triangulated is called a *triangulation* of $\mathcal{G}$. The edges of T are called *fill edges*, and for clarity of presentation, a fill edge between α and β shall often be denoted $\alpha \overset{\mathrm{f}}{—} \beta$. If there is no proper subset $T' \subset T$ such that T' is a triangulation of $\mathcal{G}$, then T is said to be a *minimal triangulation*.

A hypergraph $\mathcal{H} = (V, \mathcal{C})$ is *decomposable* if and only if $G(\mathcal{C})$ is decomposable. Any decomposable hypergraph is conformal (Lauritzen et al. 1984). If $\mathcal{C}$ is a cover of some undirected graph $\mathcal{G} = (V, E)$ and $\mathcal{H}$ is decomposable, then $\mathcal{C}$ is called a *decomposable (hypergraph) cover* of $\mathcal{G}$. $\mathcal{C}$ is a *minimal decomposable cover* of $\mathcal{G}$ if there is a minimal triangulation, T, of $\mathcal{G}$ such that $\mathcal{C}$ is the clique set of $(V, E \cup T)$.

2.2 Bayesian networks and Markov fields

For a chain graph $\mathcal{G} = (V, E)$ we consider a collection of discrete random variables $(X_\alpha)_{\alpha \in V}$ (i.e., a random field with V as index set) taking values in probability spaces $\mathrm{Sp}(X_\alpha)$. For brevity we shall interchangeably refer to $\alpha \in V$ as both a vertex and a variable. Thus we shall write e.g. α instead of X_α. For a subset $A \setminus V$ we let $\mathrm{Sp}(A) = \times_{\alpha \in A} \mathrm{Sp}(\alpha)$ (i.e., the

Cartesian product of the state spaces of the variables in A).

A probability function p is said to *factorize* according to a chain graph $\mathcal{G} = (V, E)$ if there exist non-negative functions ϕ_A defined on $\mathrm{Sp}(A)$ such that

$$p = p_V \propto \prod_{A \in \mathcal{A}} \phi_A,$$

where $\mathcal{A}$ is the set of cliques of $\mathcal{G}^m$. The functions ϕ_A are called *factor potentials* of p. If p factorizes according to $\mathcal{G}$, then $\mathcal{G}$ is said to be an *independence graph* of p, and p is a probability function of a Bayesian network with $\mathcal{G}$ as underlying graph. The independence graph captures all conditional independence statements involving pairs of sets $A, B \subseteq V$ of variables given a third set C with respect to p. We shall use the notation $A \perp\!\!\!\perp B \,|\, C$ for any triple (A, B, C) of disjoint subsets of V for which A and B are conditionally independent given C with respect to p.

If the underlying graph, $\mathcal{G} = (V, E)$, of the Bayesian network is a DAG, then all conditional independence statements captured by $\mathcal{G}$ can be found using the d-separation criterion of Pearl (1988) or the equivalent criterion of Lauritzen, Dawid, Larsen & Leimer (1990). In the general case, where $\mathcal{G}$ is a chain graph, the associated probability model is called a *block recursive model* (Lauritzen & Wermuth 1989, Højsgaard & Thiesson 1992). The Markov properties (i.e., conditional independence properties) captured by a chain graph has been described by Frydenberg (1989), and is expressed by the following theorem.

Theorem 1 (Frydenberg (1989)) *Let p factorize according to a chain graph, $\mathcal{G} = (V, E)$. Then $A \perp\!\!\!\perp B \,|\, C$ with respect to p for any subsets $A, B, C \subseteq V$ whenever C separates A from B in $(\mathcal{G}_{\mathrm{An}(A \cup B \cup C)})^m$.*

See Frydenberg (1989) for a proof. Note that the formulation of this theorem, which describes the *global chain Markov property*, is identical to the theorem of Lauritzen et al. (1990) describing the *directed global Markov property*, except that $\mathcal{G}$ is here a chain graph instead of a DAG.

Given a probability function (model) p obeying the conditional independence statements encoded by a chain graph $\mathcal{G} = (V, E)$, the corresponding collection of random variables (indexed by V) is called a *(chain) Markov field*, and p is said to be *(chain) Markov* with respect to $\mathcal{G}$. If $\mathcal{G}$ is a DAG, then the terms *directed Markov field* and *directed Markov* are often used.

2.3 Decomposable hypergraph covers and junction trees

By exploiting the conditional independence relations among the variables of a Bayesian network, the underlying joint probability space may be decomposed into a set of subspaces corresponding to a decomposable cover of the moralized network such that exact inference can be performed through a simple message passing scheme in a maximal spanning tree of the cover (Spiegelhalter 1986, Lauritzen & Spiegelhalter 1988, Jensen 1988, Jensen, Lauritzen & Olesen 1990). Technically, a decomposable cover of a Bayesian network with underlying chain graph $\mathcal{G}$ is created by triangulating $\mathcal{G}^m$. That is, the set of cliques of the triangulated graph is a decomposable cover of the network.

Jensen (1988) has shown that any maximal spanning tree of a decomposable cover, $\mathcal{C}$, can be used as the basis for a simple inward/outward message-passing scheme for propagation of evidence (belief updating) in Bayesian networks, where the nodes in such a tree contain collections of variables given as edges in the associated hypergraph (i.e., cliques in the corresponding triangulated graph) and where maximality is defined in terms of the sum of cardinalities of the intersections between adjacent nodes in the tree. Jensen (1988) named these trees *junction trees*. The nodes of a junction tree are called *(belief) universes* and the intersections between neighbouring belief universes are called *separators* (Jensen, Lauritzen & Olesen 1990). We shall henceforth refer to a junction tree by the pair $(\mathcal{C}, \mathcal{S})$ of belief universes and separators, and we shall use the terms 'belief universe' and 'clique' interchangeably. It can be shown that for each path $\langle C = C_1, \ldots, C_k = D \rangle$ in a junction tree, $C \cap D \subseteq C_i$ for all $1 \leq i \leq k$, which implies that $A \perp\!\!\!\perp B \,|\, S$ for each separator S, where A and B are the sets of variables of the two subtrees (except S) induced by the removal of the edge corresponding to S (Jensen 1988).

A junction tree $(\mathcal{C}, \mathcal{S})$ is *minimal* if there is no set $\mathcal{S}'$ such that $(\mathcal{C}, \mathcal{S}')$ is a junction tree and $\sum_{S \in \mathcal{S}'} \prod_{s \in S} \|s\| < \sum_{S \in \mathcal{S}} \prod_{s \in S} \|s\|$, where $\|s\| =$

$|\mathrm{Sp}(s)|$. The *cost*, $C(\mathcal{C})$, of a decomposable cover, $\mathcal{C}$, is given as

$$C(\mathcal{C}) = \sum_{C \in \mathcal{C}} \prod_{c \in C} \|c\| + \sum_{S \in \mathcal{S}} \prod_{s \in S} \|s\|, \tag{1}$$

where $(\mathcal{C}, \mathcal{S})$ is minimal. Obviously, to minimize the complexity of inference we should preferably choose a decomposable cover of minimum cost. Unfortunately, however, the problem of determining a minimum-cost decomposable cover is NP-hard (Wen 1990). In practice, though, simple heuristic algorithms often provide near-optimum solutions (Rose 1973, Kjærulff 1992b).

To each belief universe and each separator a *belief potential* ϕ_A is associated. The joint probability distribution, p_V, associated with a junction tree with universe set $\mathcal{C}$ and separator set $\mathcal{S}$ is proportional to the *joint (system) belief* ϕ_V given by

$$p_V \propto \phi_V = \frac{\prod_{C \in \mathcal{C}} \phi_C}{\prod_{S \in \mathcal{S}} \phi_S}.$$

A belief potential ϕ_A is *normalized* if $\sum_A \phi_A = 1$. If all belief potentials of a junction tree are normalized, then ϕ_V is normalized (i.e., $p_V = \phi_V$).

A junction tree $\Upsilon = (\mathcal{C}, \mathcal{S})$ is said to be *consistent* if

$$\sum_{C \setminus D} \phi_C \propto \sum_{D \setminus C} \phi_D \quad \text{for all } C, D \in \mathcal{C}.$$

That is, the marginal potentials for $C \cap D$ with respect to ϕ_C and ϕ_D are proportional. Consistency of Υ shall interchangeably be referred to as consistency of its associated joint belief, ϕ_V.

The basic operation of belief updating (i.e., calculation of updated potentials given incoming evidence) in junction trees is called *absorption* and is defined as follows. A belief universe C is said to *absorb* from a neighbouring universe D if the potentials ϕ_S and ϕ_C are changed to ϕ_S^* and ϕ_C^*, where $S = C \cap D$,

$$\phi_S^* = \sum_{D \setminus S} \phi_D, \quad \text{and} \quad \phi_C^* = \phi_C \frac{\phi_S^*}{\phi_S}. \tag{2}$$

For more detailed descriptions of junction trees including calculation of the initial potentials, belief updating, proofs of correctness, etc., see e.g. Jensen, Olesen & Andersen (1990), Jensen (1988), and Jensen, Lauritzen & Olesen (1990)

3 Characterization of removable edges

In this section we consider the problem of characterizing the set of edges which may be "legally" removed.

Since we consider probabilistic representations provided by decomposable covers of the moralized independence graph, the cover given by removing one of the edges must also be decomposable.

Definition 1 (Removable edges) For any decomposable cover let $\mathcal{G} = (V, E)$ be the corresponding triangulated graph. Then

$$\mathrm{rem}(\mathcal{G}) = \{e \in E \,|\, (V, E \setminus \{e\}) \text{ is triangulated}\}$$

denote the set of *removable* edges of $\mathcal{G}$. □

Thus, if $\{\alpha, \beta\} \in \mathrm{rem}(\mathcal{G})$, then removing $\alpha - \beta$ does not produce unchorded 4-cycles $\langle \alpha, \cdot, \beta, \cdot, \alpha \rangle$ or, put differently, none of the members of $\mathrm{rem}(\mathcal{G})$ are unique chords of 4-cycles. The set of removable edges is identical to the set of edges for which each member is a subset of a unique clique in the graph. These observations are formalized by

Theorem 2 *Let $\mathcal{G} = (V, E)$ be a triangulated graph with clique set $\mathcal{C}$ and intersection set $\mathcal{I} = \{C \cap D \,|\, C, D \in \mathcal{C}, C \neq D\}$, and let $E' = \mathrm{rem}(\mathcal{G})$. Let E_1, E_2 and E_3 be subsets of E given by*

(i) $E_1 = \{\{\alpha, \beta\} \in E \,|\, \mathit{adj}(\alpha) \cap \mathit{adj}(\beta) \text{ is complete}\}$
(ii) $E_2 = \{e \in E \,|\, e \text{ is member of one clique only}\}$
(iii) $E_3 = \{e \in E \,|\, \nexists I \in \mathcal{I} : e \subseteq I\}$.

E_1, E_2 and E_3 are equivalent characterizations of E', i.e. $E' \equiv E_1 \equiv E_2 \equiv E_3$.

Proof: To prove $E' \equiv E_1$ let $A = \text{adj}(\alpha) \cap \text{adj}(\beta)$ be the set of neighbours common to α and β. First assume $|A| < 2$, where $\{\alpha, \beta\} \in E_1$. Since $\mathcal{G}$ is triangulated there can be no 4-cycles including both α and β. Hence removing the edge $e = \{\alpha, \beta\}$ cannot produce unchorded 4-cycles and consequently $\mathcal{G}' = (V, E \setminus \{e\})$ is triangulated. Second, for $|A| \geq 2$ there is for each pair $\{\gamma, \delta\} \subseteq A$ of distinct common neighbours of α and β a 4-cycle $\langle \alpha, \gamma, \beta, \delta, \alpha \rangle$. If $\{\gamma, \delta\} \in E$ for all such pairs, then removing $e = \{\alpha, \beta\}$ clearly does not produce an unchorded 4-cycle and hence $\mathcal{G}'$ is triangulated. For each element $\{\alpha, \beta\}$ of the complementary set $E \setminus E_1$, $|A| \geq 2$ and there is a pair $\{\gamma, \delta\} \subseteq A$ such that $\{\gamma, \delta\} \notin E$. Hence removing the edge $e = \{\alpha, \beta\}$ creates an unchorded 4-cycle $\langle \alpha, \gamma, \beta, \delta, \alpha \rangle$ rendering $\mathcal{G}' = (V, E \setminus \{e\})$ non-triangulated. Thus $E_1 \equiv E'$.

To prove $E_1 \equiv E_2$ assume $\{\alpha, \beta\} \in E_1$. Then $A = \text{adj}(\alpha) \cap \text{adj}(\beta)$ induces a complete subgraph of $\mathcal{G}$. $C = A \cup \{\alpha, \beta\}$ induces a clique; otherwise there would be a clique $C' \supset C$ such that $C' \setminus C \subset A$ implying $C' = C$. Now, assume there is another clique C^* such that $\{\alpha, \beta\} \subset C^*$. But as $\alpha — \beta$ is not a unique chord of a 4-cycle, $A \cup C^*$ must induce a complete subgraph. This, however, contradicts the fact that C is a clique. Thus C uniquely contains $\{\alpha, \beta\}$.

Alternatively, E_3 may be given as $\{\{\alpha, \beta\} \in E | \not\exists C_1, C_2 \in \mathcal{C} : \{\alpha, \beta\} \subseteq C_1, \{\alpha, \beta\} \subseteq C_2\}$. Thus for each element $e \in E_3$ there is a unique clique $C \in \mathcal{C}$ such that $e \subseteq C$ and hence $E_2 \equiv E_3$. □

From Theorem 2 it follows that $\text{rem}(\mathcal{G}) \cap T = \emptyset$, where T is a minimal triangulation of (V, E) and $\mathcal{G} = (V, E \cup T)$. Thus in a minimally triangulated graph the set of removable edges is a subset of the edges of the original moral graph.

An important observation to make at this point is that any triangulated graph contains at least one removable edge (provided disconnecting the graph is allowed). In order to formally prove this statement, we first prove the following two lemmas.

Lemma 1 *If a graph $\mathcal{G} = (V, E)$ is triangulated, then there is a decomposition of $\mathcal{G}$.*

Proof: Follows immediately from the fact that a graph is triangulated if and only if it is decomposable (Lauritzen et al. 1984). □

Lemma 2 (Dirac (1961)) *Let $\mathcal{G} = (V, E)$ be a triangulated graph with $|V| \geq 2$. Then $\mathcal{G}$ has at least two simplicial vertices. If $\mathcal{G}$ is not complete, these can be chosen to be non-adjacent.*

Proof: The proof is by induction on $|V|$. For $|V| = 2$ the result is obvious. Assume that the result holds for all $|V| \leq k$ and let $|V| = k+1$. If $\mathcal{G}$ is complete, the result is vacuously true. Otherwise, by Lemma 1, there exists a decomposition (A, B, C) of $\mathcal{G}$. If $A \cup C$ is complete, then $\mathcal{G}_{A \cup C}$ contains at least two simplicial vertices since A and C are disjoint and non-empty. At least one of these, say α, must be in A. If $A \cup C$ is not complete, then the induction hypothesis used on $\mathcal{G}_{A \cup C}$ yields a pair (α, β) of non-adjacent vertices that are simplicial in $\mathcal{G}_{A \cup C}$. Since C is complete, at least one of these, say α, must be in A. By symmetry there is a vertex $\gamma \in B$ that is simplicial in $\mathcal{G}_{B \cup C}$. Since C separates A from B, (α, γ) must be a pair of non-adjacent simplicial vertices in $\mathcal{G}$. □

Theorem 3 *Let $\mathcal{G} = (V, E)$ be a triangulated graph. Then there is at least one edge $e \in E$ such that $\mathcal{G}' = (V, E \setminus \{e\})$ is also triangulated.*

Proof: According to Lemma 2, $\mathcal{G}$ has at least two simplicial vertices. Let α be a simplicial vertex of $\mathcal{G}$, and let E_α be the set of edges incident to α. As $\text{adj}(\alpha)$ induces a complete subgraph of $\mathcal{G}$, any subset $E'_\alpha \subseteq E_\alpha$ can be removed without compromising the chordality of $\mathcal{G}' = (V, E \setminus E'_\alpha)$. □

4 Consequences of edge removal

As conceived in this paper, any Bayesian network model with underlying probability model, p, consists of three distinct components:

1. a potential representation of p,
2. a decomposable hypergraph cover, and
3. an independence graph of p,

where, as discussed in Section 2.3, the potential representation is given as a set of belief potentials defined on the belief universes and separators of a junction tree given by a decomposable hypergraph cover of the (moralized) independence graph.

We shall assume that the decomposable covers considered are minimal. Thus, only moral edges or edges of the independence graph may be removed. Therefore, the removal of an edge manifests itself in modifying all of the three above-mentioned components of the Bayesian model. Below we discuss these three kinds of modifications separately.

4.1 Potential representation

First we show that only the belief potential of the universe exposed to approximation by removing an edge gets affected. A straightforward consequence thereof, is that the approximated joint belief is consistent if the joint belief before the removal was consistent.

Lemma 3 *For the belief potentials ϕ_C and ψ_C of a belief universe C, where*

$$\psi_C \propto \frac{\sum_\alpha \phi_C \sum_\beta \phi_C}{\sum_{\alpha,\beta} \phi_C} \quad \Leftrightarrow \quad \alpha \perp\!\!\!\perp \beta \,|\, C \setminus \{\alpha, \beta\} \text{ with respect to } \psi. \tag{3}$$

and $\{\alpha, \beta\} \subseteq C$, the following relations hold:

$$\sum_\alpha \phi_C \propto \sum_\alpha \psi_C, \tag{4}$$

$$\sum_\beta \phi_C \propto \sum_\beta \psi_C. \tag{5}$$

Proof: By definition

$$\sum_\alpha \psi_C \propto \sum_\alpha \frac{\sum_\alpha \phi_C \sum_\beta \phi_C}{\sum_{\alpha,\beta} \phi_C} = \frac{\sum_\alpha \phi_C \sum_{\alpha,\beta} \phi_C}{\sum_{\alpha,\beta} \phi_C} = \sum_\alpha \phi_C.$$

Since (3) is symmetric in α and β, (5) follows from (4). □

Thus, if C is the unique universe containing $\{\alpha, \beta\}$ and $C_1, \ldots, C_k$ are the neighbouring universes of C with $S_1 \ldots, S_k$ being the associated separators, then the removal of $\alpha — \beta$ does not affect the separator potentials $\phi_{S_1} \ldots, \phi_{S_k}$, since $S_i \subseteq C \setminus \{\alpha\}$ or $S_i \subseteq C \setminus \{\beta\}$ for all $1 \leq i \leq k$. Therefore we easily get the following corollary.

Corollary 1 *If ϕ is a consistent joint belief of a junction tree $\Upsilon = (\mathcal{C}, \mathcal{S})$, where $C \in \mathcal{C}$ is the unique universe containing $\{\alpha, \beta\}$ and*

$$\psi_{C'} \propto \begin{cases} \dfrac{\sum_\alpha \phi_C \sum_\beta \phi_C}{\sum_{\alpha,\beta} \phi_C} & \text{if } C' = C, \\ \phi_{C'} & \text{otherwise,} \end{cases} \tag{6}$$

where $C' \in \mathcal{C}$, then ψ is also a consistent joint belief of Υ.

Proof: Let $\mathrm{adj}(C) = \{C_1, \ldots, C_k\}$ and let $S_i = C \cap C_i$ for all $i = 1, \ldots, k$. Since C is the unique universe containing $\{\alpha, \beta\}$ either $S_i \subseteq C \setminus \{\alpha\}$ or $S_i \subseteq C \setminus \{\beta\}$ for all $1 \leq i \leq k$. Hence by Lemma 3, $\phi_{S_i} \propto \psi_{S_i}$ and the result follows. □

4.2 Junction tree

When an edge $\{\alpha, \beta\} \subseteq C \in \mathcal{C}$ is removed from a triangulated graph $\mathcal{G} = (V, E)$ with clique set $\mathcal{C}$ one of the following three mutually exclusive statements must hold for the complete sets $C_\alpha = C \setminus \{\beta\}$ and $C_\beta = C \setminus \{\alpha\}$.

(i) Both C_α and C_β are cliques of $\mathcal{G}' = (V, E \setminus \{\{\alpha, \beta\}\})$, or

(ii) C_α (or C_β) is a clique of $\mathcal{G}'$ and C_β (or C_α) is not, or

(iii) neither C_α nor C_β are cliques of $\mathcal{G}'$.

To see that all three cases can occur consider the examples of Figure 1. The cost reduction (i.e., reduction of the sum of the state space sizes of the belief universes — cf. (1)) achieved by the removal is

$$\begin{aligned} \sigma(\alpha, \beta) &= \|C\| - (I_{\mathcal{C}^-}(C_\alpha)\|C_\alpha\| + I_{\mathcal{C}^-}(C_\beta)\|C_\beta\|) \\ &= \|C\| \left(1 - I_{\mathcal{C}^-}(C_\alpha)\|\beta\|^{-1} - I_{\mathcal{C}^-}(C_\beta)\|\alpha\|^{-1}\right), \end{aligned} \tag{7}$$

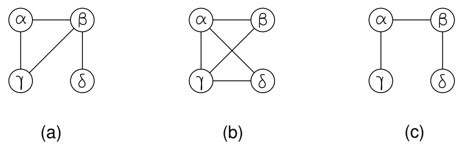


Figure 1: Removal of α—β. In part (a), both $\{\alpha, \gamma\}$ and $\{\beta, \gamma\}$ are new cliques. In part (b), $\{\beta, \gamma\}$ is a new clique, whereas $\{\alpha, \gamma\}$ is not. In part (c), neither $\{\alpha\}$ nor $\{\beta\}$ are new cliques.

where $\mathcal{C}^- = \mathcal{C} \setminus \{C\}$ and I is an indicator function given by

$$I_Z(z) = \begin{cases} 0 & \text{if } \exists z' \in Z : z \subseteq z' \\ 1 & \text{otherwise.} \end{cases}$$

Note that $0 \leq \sigma \leq \|C\|$, where $\sigma = 0$ when $\|\alpha\| = \|\beta\| = 2$ and $I_{\mathcal{C}^-}(C_\alpha) = I_{\mathcal{C}^-}(C_\beta) = 1$, and $\sigma = \|C\|$ when $I_{\mathcal{C}^-}(C_\alpha) = I_{\mathcal{C}^-}(C_\beta) = 0$.

Let $C_1, \ldots, C_k$ be the neighbours of C in a junction tree $\Upsilon = (\mathcal{C}, \mathcal{S})$ corresponding to $\mathcal{G}$, let $S_1, \ldots, S_k$ be the associated separators, and let ϕ be the joint belief of Υ. The new junction tree $\Upsilon' = (\mathcal{C}', \mathcal{S}')$ corresponding to $\mathcal{G}'$ has joint belief ψ and is constructed as follows, where the procedures are described according to the three above-mentioned situations.

Re (i) When C_α and C_β are both cliques of $\mathcal{G}'$ we let $\mathcal{C}' = \mathcal{C}^- \cup \{C_\alpha, C_\beta\}$ and $\mathcal{S}' = \mathcal{S} \cup \{S\}$, where $S = C_\alpha \cap C_\beta = C \setminus \{\alpha, \beta\}$, and connect C_i $(1 \leq i \leq k)$ to either C_α or C_β (note that S_i is a subset of either C_α or C_β for all $i = 1, \ldots, k$). To be precise,

1. Remove C from Υ.
2. Create a separator $S = C \setminus \{\alpha, \beta\}$ and let $\psi_S = \phi_S = \sum_{\alpha,\beta} \phi_C$.
3. Create two belief universes on C_α and C_β, respectively, and let $\psi_{C_\alpha} = \phi_{C_\alpha} = \sum_\beta \phi_C$ and $\psi_{C_\beta} = \phi_{C_\beta} = \sum_\alpha \phi_C$.
4. Connect C_α and C_β via S.

5. For $i = 1$ to k do
   if $\alpha \in S_i$ or ($\beta \notin S_i$ and $\|C_\alpha\| < \|C_\beta\|$) then
     connect C_i and C_α via S_i
   else
     connect C_i and C_β via S_i.

Since $S_i \cap S_j \subseteq S$ for all $1 \leq i \neq j \leq k$, Υ' is clearly a junction tree. Figure 2(i) shows Υ', where new objects and links are shown in bold.

Re (ii) Assume that C_α is a clique of $\mathcal{G}'$ (the situation is symmetric if C_β is a clique of $\mathcal{G}'$). Thus, $\mathcal{C}' = \mathcal{C}^- \cup \{C_\alpha\}$. Without loss of generality, assume that $C_\beta \subset C_k$ (i.e., $C_\beta = S_k$) such that $\|C_k\| \leq \|C_i\|$ for all $C_i \supset C_\beta$. Therefore, we connect C_α and C_k via a new separator $S = C_\alpha \cap C_k = C \setminus \{\alpha, \beta\}$ and repeat the above procedure with C_β replaced by C_k:

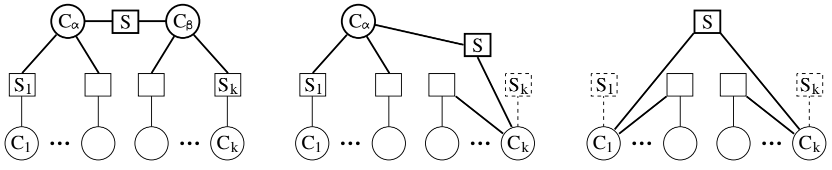
1. Remove C and S_k from Υ.
2. Create a separator $S = C \setminus \{\alpha, \beta\}$ and let $\psi_S = \phi_S = \sum_{\alpha,\beta} \phi_C$.
3. Create a belief universe on C_α and let $\psi_{C_\alpha} = \phi_{C_\alpha} = \sum_\beta \phi_C$.
4. Connect C_α and C_k via S.
5. For $i = 1$ to $k - 1$ do
   if $\alpha \in S_i$ or ($\beta \notin S_i$ and $\|C_\alpha\| < \|C_k\|$) then
     connect C_i and C_α via S_i
   else
     connect C_i and C_k via S_i.

Again, Υ' is clearly a junction tree, since $S_i \cap S_j \subseteq S$ for all $1 \leq i \neq j \leq k - 1$. Figure 2(ii) shows Υ'.

Re (iii) Finally, when neither C_α nor C_β are cliques of $\mathcal{G}'$, assume (without loss of generality) that $C_\alpha \subset C_1$ and $C_\beta \subset C_k$ (i.e., $C_\alpha = S_1$ and $C_\beta = S_k$) such that $\|C_1\| \leq \|C_i\|$ for all $C_i \supset C_\alpha$ and $\|C_k\| \leq \|C_j\|$ for all $C_j \supset C_\beta$. Thus, we connect C_1 and C_k via a new separator $S = C_1 \cap C_k = C \setminus \{\alpha, \beta\}$ and use the procedure of case (i) with C_α replaced by C_1 and C_β replaced by C_k:

1. Remove C, S_1, and S_k from Υ.
2. Create a separator $S = C \setminus \{\alpha, \beta\}$ and let $\psi_S = \phi_S = \sum_{\alpha,\beta} \phi_C$.
3. Connect C_1 and C_k via S.
4. For $i = 2$ to $k-1$ do
   if $\alpha \in S_i$ or ($\beta \notin S_i$ and $\|C_1\| < \|C_k\|$) then
   connect C_i and C_1 via S_i
   else
   connect C_i and C_k via S_i.

Figure 2(iii) shows Υ'.



(i) (ii) (iii)

Figure 2: Removal of $\alpha \,—\, \beta$ results in a junction tree, Υ', with a new separator $S = C \setminus \{\alpha, \beta\}$ such that the removal of S produces a subtree Υ'_1 containing α but not β and a subtree Υ'_2 containing β but not α.

Note that in all three cases we have $S = C \setminus \{\alpha, \beta\}$ meaning that S separates Υ' into two subtrees $\Upsilon'_A = (\mathcal{C}'_A, \mathcal{S}'_A)$ and $\Upsilon'_B = (\mathcal{C}'_B, \mathcal{S}'_B)$, where A and B are the corresponding sets of variables such that $\alpha \in A$ and $\beta \in B$. From Lemma 3 it follows trivially that

$$\psi_{A \cup S} = \frac{\prod_{C' \in \mathcal{C}'_A} \psi_{C'}}{\prod_{S' \in \mathcal{S}'_A} \psi_{S'}} = \frac{\prod_{C' \in \mathcal{C}'_A} \phi_{C'}}{\prod_{S' \in \mathcal{S}'_A} \phi_{S'}} = \phi_{A \cup S}$$

and similarly $\psi_{B \cup S} = \phi_{B \cup S}$. Therefore,

$$\psi = \frac{\phi_{A \cup S} \phi_{B \cup S}}{\phi_S}. \tag{8}$$

4.3 Independence graph

Since $\psi_A = \psi_{A \cup S} = \phi_{A \cup S}$, the independence relations among the variables of the set A are unaltered by the removal of $\alpha \,—\, \beta$, where ϕ is the 'correct' joint belief and ψ is the approximate one given by (8). The same applies to the sets B and S. That is, the marginal independence graphs for A, B and S are identical for ϕ and ψ. Therefore, the problem of determining the independence graph of ψ may be formulated as the problem of combining marginal independence graphs such that the independence statements expressed by these are not violated and such that the combined graph represents the fact that $A \perp\!\!\!\perp B \,|\, S$ (or $A \setminus S \perp\!\!\!\perp B \setminus S \,|\, S$ to be exact).

Given an independence graph of a probability function (belief potential), $p = p_V$, we need a way of establishing an independence graph of any marginal p_A, $A \subseteq V$. First, however, we shall introduce the notion of minimal directedness.

Definition 2 (Minimal directedness) A chain graph $\mathcal{G} = (V, E)$ is *minimally directed* if it does not contain chain components $K, P \subset V$ such that $\mathrm{pa}(K)$ is a non-empty, complete subset of P. □

Theorem 4 *Let $\mathcal{G} = (V, E)$ be a chain graph containing chain components $K, P \subset V$ such that $\mathrm{pa}(K)$ is a non-empty, complete subset of P (i.e., $\mathcal{G}$ is not minimally directed) and let the graph $\mathcal{G}'$ be obtained by replacing all directed edges between K and P by undirected edges. Then $\mathcal{G}'$ is a chain graph and the independence properties represented by $\mathcal{G}$ and $\mathcal{G}'$ are identical.*

Proof: The addition of (β, α) to E for each $\alpha \to \beta$ ($\beta \in K, \alpha \in \mathrm{pa}(\beta)$) possibly introduces weakly directed cycles $\langle \beta, \alpha = \alpha_0, \ldots, \alpha_m = \gamma, \delta = \beta_0, \ldots, \beta_n = \beta \rangle$, where $\{\alpha_i\}_{i=0}^m \subseteq P$ and $\{\beta_i\}_{i=0}^n \subseteq K$. But, since all involved edges are undirected, no (strongly) directed cycles are introduced, and hence $\mathcal{G}'$ is a chain graph.

To see that $\mathcal{G}$ and $\mathcal{G}'$ represent identical independence statements, recall the criterion of Theorem 1 stating that $A \perp\!\!\!\perp B \,|\, C$ whenever C separates A from B in $(\mathcal{G}_{\mathrm{An}(A \cup B \cup C)})^m$. When $\{\alpha, \beta\} \cap \mathrm{An}(A \cup B \cup C) = \emptyset$ in $\mathcal{G}$ (and $\mathcal{G}'$) for all $\beta \in K$ and $\alpha \in \mathrm{pa}(\beta)$, $(\mathcal{G}_{\mathrm{An}(A \cup B \cup C)})^m$ and

$(\mathcal{G}'_{\mathrm{An}(A\cup B\cup C)})^m$ are obviously identical. When $\{\alpha,\beta\} \cap \mathrm{An}(A \cup B \cup C) \neq \emptyset$ in $\mathcal{G}$ for some $\beta \in K$ and $\alpha \in \mathrm{pa}(\beta)$, $(\mathcal{G}_{\mathrm{An}(A\cup B\cup C)})^m$ and $(\mathcal{G}'_{\mathrm{An}(A\cup B\cup C)})^m$ are identical with the possible exception of a subgraph $\mathcal{G}'_{K^*}$, where $\mathrm{an}(\alpha) \supset K^* \subset K$, when $(A \cup B \cup C) \cap K^* = \emptyset$. But in that case $\mathcal{G}'_{K^*}$ has no influence on the separability of A and B by C, since K^* and $\mathrm{An}(A \cup B \cup C) \setminus K^*$ are separated by the complete set $\mathrm{pa}(\alpha)$ in $(\mathcal{G}'_{\mathrm{An}(A\cup B\cup C)})^m$. □

Thus, for any independence graph which is not minimally directed, there is a minimally directed graph with identical independence structure obtained by successive replacements of sets of directed edges with undirected edges, where the set of edges to be replaced complies with the restrictions of Theorem 4. Frydenberg (1989) shows an equivalent result stating that two chain graphs $\mathcal{G} = (V,E)$ and $\mathcal{G}' = (V,E')$ with identical undirected graphs (i.e., graphs obtained by replacing all directed edges with undirected ones) represent the same independence statements if $E \subseteq E'$ (i.e., $\mathcal{G}$ may contain directed edges where the equivalent ones in $\mathcal{G}'$ are undirected) and $\mathrm{pa}(K)$ is complete in $(\mathcal{G}_L)^m$ for each chain component K of $\mathcal{G}'$, where $L = K \cup \mathrm{pa}(K)$.

Definition 3 (Correct and perfect independence graphs) Let $\mathcal{G} = (V,E)$ be any chain graph and p any probability function (belief potential) defined on $\mathrm{Sp}(V)$. If $\mathcal{G}$ does not represent (conditional) independence properties inconsistent with p, then $\mathcal{G}$ is a *correct* independence graph of p. If $\mathcal{G}$ is a correct independence graph of p and all (conditional) independence properties of p are represented by $\mathcal{G}$, then $\mathcal{G}$ is a *perfect* independence graph of p. □

We shall often say '$\mathcal{G}$ is an independence graph of p' with correctness being understood. Further, we shall use the shorthands: '$\mathcal{G}$ is correct' and '$\mathcal{G}$ is perfect', when it is understood that $\mathcal{G}$ is an independence graph of a given probability function.

Below, we use the symbol $\otimes$ to denote the binary operator producing the set of all ordered pairs of distinct elements of its arguments.

Definition 4 (Marginal independence graph) Let $\mathcal{G} = (V,E)$ be any chain graph and let $\alpha \in V$ be any vertex of $\mathcal{G}$. Then $\mathcal{G}^{\downarrow}_{V\setminus\{\alpha\}} = (V \setminus \{\alpha\}, E^{\downarrow}_{V\setminus\{\alpha\}})$ is given by $E^{\downarrow}_{V\setminus\{\alpha\}} = E' \cup E^*$, where

$$E' = E \setminus E_\alpha \cup (\mathrm{nb}(\alpha) \otimes \mathrm{nb}(\alpha)) \cup (\mathrm{pa}(\alpha) \times (\mathrm{nb}(\alpha) \cup \mathrm{ch}(\alpha))) \cup (\mathrm{nb}(\alpha) \times \mathrm{ch}(\alpha)),$$

(E_α is the set of edges incident to α) and E^* denotes the set of edges rendering $\mathrm{ch}(\alpha)$ complete in such a way that

$\beta — \gamma$ if β and γ belong to the same chain component,

$\beta \rightarrow \gamma$ if the chain component containing β precedes the component containing γ,

$\beta \leftarrow \gamma$ otherwise,

for any two non-adjacent vertices $\beta, \gamma \in \mathrm{ch}(\alpha)$ in $\mathcal{G}$. □

In words, $\mathcal{G}^{\downarrow}_{V\setminus\{\alpha\}}$ is constructed by making the neighbours of α complete, adding directed edges from the parents of α to its neighbours and children, adding directed edges from the neighbours of α to its children, making the children of α complete in such a way that no (strongly) directed paths are introduced, and removing α and the edges incident to it. Figure 3 illustrates this process.

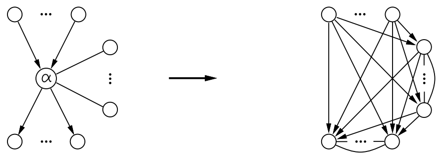


Figure 3: Marginalization with respect to an unobserved variable α render $\mathrm{ch}(\alpha)$ and $\mathrm{nb}(\alpha)$ complete and adds directed edges from $\mathrm{pa}(\alpha)$ to $\mathrm{nb}(\alpha) \cup \mathrm{ch}(\alpha)$ and from $\mathrm{nb}(\alpha)$ to $\mathrm{ch}(\alpha)$.

Lemma 4 *For any chain graph $\mathcal{G} = (V, E)$ and any $\alpha \in V$, $\mathcal{G}^{\downarrow}_{V\setminus\{\alpha\}}$ is a chain graph.*

Proof: Since $\mathrm{nb}(\alpha)$ is a subset of a single chain component in $\mathcal{G}$, completion of $\mathrm{nb}(\alpha)$ by addition of undirected edges cannot introduce directed cycles. The addition of $\pi \to \mu$ for each $\pi \in \mathrm{pa}(\alpha)$ and $\mu \in \mathrm{nb}(\alpha) \cup \mathrm{ch}(\alpha)$ cannot introduce directed cycles either, since π belongs to a chain component of lower order than the component containing μ (i.e., any path from π to μ is strongly directed). The same is obviously true for the directed edges from $\mathrm{nb}(\alpha)$ to $\mathrm{ch}(\alpha)$. Finally, since the edges added between pairs $\beta, \gamma \in \mathrm{ch}(\alpha)$ are directed only if β and γ belong to different chain components and such that the direction of the edge follows the ordering of the components, we conclude that $\mathcal{G}^{\downarrow}_{V\setminus\{\alpha\}}$ is a chain graph. □

Below we prove that $\mathcal{G}^{\downarrow}_{V\setminus\{\alpha\}}$ is an independence graph of the marginal distribution $\sum_\alpha p$ if $\mathcal{G}$ is an independence graph of a probability function p. Note that correctness of $\mathcal{G}' = \mathcal{G}^{\downarrow}_{V\setminus\{\alpha\}}$ follows if separation of A and B by C in $(\mathcal{G}'_{\mathrm{An}(A\cup B\cup C)})^m$ implies separation in $(\mathcal{G}_{\mathrm{An}(A\cup B\cup C)})^m$ as well, and that perfectness of $\mathcal{G}^{\downarrow}_{V\setminus\{\alpha\}}$ follows if separation in $(\mathcal{G}_{\mathrm{An}(A\cup B\cup C)})^m$ implies separation in $(\mathcal{G}'_{\mathrm{An}(A\cup B\cup C)})^m$ provided $\mathcal{G}$ is perfect.

Note: If a path between A and B contains one or more vertices in C we shall say that the path is *blocked* (by C). Similarly, if a path between A and B contains no vertices in C, it is *unblocked*.

Theorem 5 *Let the chain graph $\mathcal{G} = (V, E)$ be an independence graph of $p = p_V$, where $\alpha \in V$ is any unobserved variable. Then $\mathcal{G}^{\downarrow}_{V\setminus\{\alpha\}}$ is an independence graph of $p_{V\setminus\{\alpha\}} = \sum_\alpha p_V$.*

Proof: Let $A, B, C \subseteq V \setminus \{\alpha\}$ be any disjoint subsets such that C separates A and B in $(\mathcal{G}'_{\mathrm{An}(A\cup B\cup C)})^m$, where $\mathcal{G}' = \mathcal{G}^{\downarrow}_{V\setminus\{\alpha\}}$, and let $\pi \in A$ and $\mu \in B$ be arbitrary vertices. Assume that there is an unblocked path between π and μ in $(\mathcal{G}_{\mathrm{An}(A\cup B\cup C)})^m$. Since $\mathcal{G}$ is a subgraph of $(V, E^{\downarrow}_{V\setminus\{\alpha\}} \cup E_\alpha)$, the path must include α. Let $\langle \pi, \dots, \beta, \alpha, \gamma, \dots, \mu \rangle$ be such a path. Again, since $\mathcal{G}$ is a subgraph of $(V, E^{\downarrow}_{V\setminus\{\alpha\}} \cup E_\alpha)$, the

subpaths $\langle \pi, \dots, \beta \rangle$ and $\langle \gamma, \dots, \mu \rangle$ must also be paths in $(\mathcal{G}'_{\mathrm{An}(A\cup B\cup C)})^m$. By the construction of $E^{\downarrow}_{V\setminus\{\alpha\}}$, $\beta \sim \gamma$ in $(\mathcal{G}'_{\mathrm{An}(A\cup B\cup C)})^m$ violating the assumption of A and B being separated by C in $(\mathcal{G}'_{\mathrm{An}(A\cup B\cup C)})^m$. □

The following example shows that $\mathcal{G}^{\downarrow}_{V\setminus\{\alpha\}}$ is not necessarily perfect when $\mathcal{G}$ is perfect. Let $V = \{\alpha, \beta, \gamma, \delta, \epsilon\}$ and $p = p_V$ be Markov with respect to the DAG in Figure 4a. Since $\beta \not\perp\!\!\!\perp \delta \,|\, \{\gamma, \epsilon\}$ with respect to

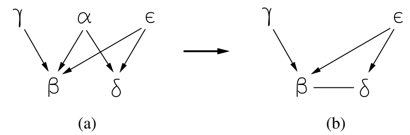


Figure 4: $\gamma \perp\!\!\!\perp \epsilon \,|\, \delta$ with respect to to p (and $\sum_\alpha p$) which is Markov with respect to the DAG in part (a). However, $\sum_\alpha p$ is not Markov with respect to the graph in part (b), since according to that $\gamma \not\perp\!\!\!\perp \epsilon \,|\, \delta$.

to p (and $p_{V\setminus\{\alpha\}}$), we must have $\beta \sim \delta$ in an independence graph of $p_{V\setminus\{\alpha\}} = \sum_\alpha p$, and, since $\gamma \perp\!\!\!\perp \epsilon$ and $\gamma \not\perp\!\!\!\perp \epsilon \,|\, \beta$, a candidate independence graph of $p_{V\setminus\{\alpha\}}$ could be the one in Figure 4b. However, since $\gamma \perp\!\!\!\perp \epsilon \,|\, \delta$ with respect to to p (and $p_{V\setminus\{\alpha\}}$), this graph is not perfect, but it is correct, since it does not represent non-existing conditional independences. Thus, all the independence properties of $p_{V\setminus\{\alpha\}}$ cannot be represented by a single chain graph. If we want a perfect representation, a more sophisticated language must be adopted. One such language may be given by the class of annotated graphs (Geva & Paz 1992). However, in the present paper we shall refrain from pursuing this any further.

Corollary 2 *If $\mathcal{G} = (V, E)$ is a chain graph and $A \subseteq V$ is a set of founders, then $\mathcal{G}^{\downarrow}_A = \mathcal{G}_A$.*

Proof: Follows immediately from Definition 4. □

Corollary 3 *Let $\mathcal{G} = (V, E)$ be a chain graph and $\alpha, \beta \in V$ any pair of non-adjacent vertices of $\mathcal{G}$. Then $\alpha \sim \beta$ in $\mathcal{G}^{\downarrow}_{V \setminus \mathrm{an}(\{\alpha,\beta\})}$ if $an(\alpha) \cap an(\beta) \neq \emptyset$.*

Proof: The proof is by induction on $|\mathrm{an}(\{\alpha, \beta\})|$. If $\mathrm{an}(\{\alpha, \beta\}) = \{\gamma\}$, then $\mathrm{an}(\alpha) = \mathrm{an}(\beta) = \{\gamma\}$ and, by Definition 4, $\alpha \sim \beta$ in $\mathcal{G}^{\downarrow}_{V\setminus\{\gamma\}} = \mathcal{G}^{\downarrow}_{V\setminus \mathrm{an}(\{\alpha,\beta\})}$. Assume that the result holds for $|\mathrm{an}(\{\alpha, \beta\})| < k$, and let $|\mathrm{an}(\{\alpha, \beta\})| = k$. Let γ be the variable to be eliminated first and let $A = (\mathrm{ch}(\gamma) \cup \mathrm{nb}(\gamma)) \cap \mathrm{an}(\{\alpha, \beta\})$.

First, assume that $\gamma \in \mathrm{an}(\alpha) \cap \mathrm{an}(\beta)$. If $A = \{\delta\}$, then $\delta \in \mathrm{an}(\alpha) \cap \mathrm{an}(\beta)$ in $\mathcal{G}^{\downarrow}_{V\setminus\{\gamma\}}$, and, by the induction hypothesis, the result holds. Let $\mu, \pi \in A$ be any two distinct vertices in A. Since $\mu \sim \pi$ in $\mathcal{G}^{\downarrow}_{V\setminus\{\gamma\}}$, at least one of μ, π will be a member of $\mathrm{an}(\alpha) \cap \mathrm{an}(\beta)$ in $\mathcal{G}^{\downarrow}_{V\setminus\{\gamma\}}$ making the result hold.

Second, assume that $\gamma \in \mathrm{an}(\alpha)$ and let $\delta \in \mathrm{an}(\alpha) \cap \mathrm{an}(\beta)$. If $\delta \notin \mathrm{an}(\gamma)$, then the elimination of γ clearly does not affect δ's membership of $\mathrm{an}(\alpha) \cap \mathrm{an}(\beta)$. If $\delta \in \mathrm{an}(\gamma)$, then, by Definition 4, the elimination of γ will add edges such that δ's membership of $\mathrm{an}(\alpha) \cap \mathrm{an}(\beta)$ will not be affected. In both cases, the induction hypothesis makes the result hold. □

Definition 4 provides a method for constructing a (marginal) independence graph of the marginal distribution $p_{V\setminus\{\alpha\}}$ (i.e., it deals with the case where α is unobserved). However, the construction of an independence graph of the approximate joint belief $\psi = \phi_{A\cup S}\phi_{B\,|\,S}$ involves combination of a marginal independence graph and a conditional (marginal) independence graph. The (conditional) independence graph of the conditional distribution $p_{V\,|\,\alpha}$ is obtained simply by moralizing the subgraph induced by $\mathrm{An}(\alpha)$ and removing α and the edges incident to it.

Theorem 6 *Let the chain graph $\mathcal{G} = (V, E)$ be an independence graph of $p = p_V$ and let $E^m_{\mathrm{An}(A)}$ be the edge set of $(\mathcal{G}_{\mathrm{An}(A)})^m$. Then $\mathcal{G}' = (V, E \cup E^m_{\mathrm{An}(A)})$ is a chain graph and a conditional independence graph of $p_{V\,|\,A}$.*

Proof: Since $\mathcal{G}$ is a chain graph, $\mathrm{An}(A)$ is a chain component of $\mathcal{G}'$, and there are no directed edges from $V \setminus \mathrm{An}(S)$ to $\mathrm{An}(S)$, $\mathcal{G}'$ is also a chain graph. The fact that $\mathcal{G}'$ is a conditional independence graph of $p_{V\,|\,A}$ follows directly from Theorem 1. □

By the methods of Definition 4 and Theorem 6 we can construct any marginal independence graph (possibly conditional on a set of variables) by successive removal of the relevant variables.

Note that the presence of the set A and the edges incident to A in the independence graph of $p_{V\,|\,A}$ is unnecessary for a correct interpretation of the conditional independence relations among variables in $V \setminus A$ given A (this is stated formally by Corollary 4 below). However, when combining a conditional and a marginal independence graph to obtain a joint independence graph, these vertices and edges are needed. Thus, we shall assume that the variables upon which we condition are included in the conditional independence graphs considered. Theorem 7 below states that a joint independence graph can be formed by simple graph union of a conditional and a marginal independence graph.

Corollary 4 *Let the chain graph $\mathcal{G}_{V\,|\,A} = (V, E)$ be a conditional independence graph of $p_{V\,|\,A}$, where $A \subseteq V$. Then $\mathcal{G}'_{V\,|\,A} = (V, E \setminus (A \otimes A) \cup (A \times nb(A)))$ is a chain graph and a conditional independence graph of $p_{V\,|\,A}$.*

Proof: Obviously, the operation does not introduce directed cycles; $\mathcal{G}'_{V\,|\,A}$ is therefore a chain graph. Any path involving a variable in A is blocked. Hence, the removal of any subset, $\{(\alpha, \beta)\} \subseteq E$, where $\alpha, \beta \in A$ or $\alpha \in \mathrm{nb}(A)$ and $\beta \in A$, does not alter the conditional independence representation given A. □

When combining a conditional and a marginal independence graph, we shall assume that the conditional one has been exposed to the transformation described in Corollary 4, since, in general, we cannot otherwise be sure that the resulting graph will be a chain graph.

Theorem 7 *Let the chain graph $\mathcal{G}^{\downarrow}_{A\cup S} = (A, E^{\downarrow}_A)$ be a marginal independence graph of $p_{A\cup S}$ and the chain graph $\mathcal{G}^{\downarrow}_{B\,|\,S} = (B, E^{\downarrow}_B)$ a* *conditional independence graph of $p_{B\,|\,S}$, where $A, B, S \subseteq V$ such that $A \cap B = S$ and $A \perp\!\!\!\perp B\,|\,S$ with respect to $p = p_V$. Then $\mathcal{G}' = \mathcal{G}^{\downarrow}_{A\cup S} \cup \mathcal{G}^{\downarrow}_{B\,|\,S}$ is a chain graph and an independence graph of p. Further, $\mathcal{G}'$ is perfect if both $\mathcal{G}^{\downarrow}_{A\cup S}$ and $\mathcal{G}^{\downarrow}_{B\,|\,S}$ are perfect.*

Proof: For $\mathcal{G}'$ to be an independence graph of p we must require that

(i) any path from $\alpha \in A \setminus S$ to $\beta \in B \setminus S$ is blocked by S in $(\mathcal{G}'_{\mathrm{An}(\{\alpha,\beta\}\cup S)})^m$ (i.e., the fact that $A \perp\!\!\!\perp B\,|\,S$ with respect to p must be reflected by $\mathcal{G}'$), and

(ii) $\mathcal{G}'^{\downarrow}_{A\cup S}$ and $\mathcal{G}'^{\downarrow}_{B\cup S}$ are independence graphs of $p_{A\cup S}$ and $p_{B\cup S}$, respectively.

Since S is a founder set of $\mathcal{G}'_{B\cup S}$, there can be no moral edge between α and β in $(\mathcal{G}'_{\mathrm{An}(\{\alpha,\beta\}\cup S)})^m$ whenever $\alpha \in A \setminus S$ and $\beta \in B \setminus S$. Thus, since $A \cap B = S$ and $\mathcal{G}' = \mathcal{G}^{\downarrow}_{A\cup S} \cup \mathcal{G}^{\downarrow}_{B\,|\,S}$, (i) is fulfilled.

Now, there can be no directed cycles in $\mathcal{G}'$, since (i) is fulfilled, S is a founder set of $\mathcal{G}'_{B\cup S} = \mathcal{G}^{\downarrow}_{B\,|\,S}$, and both $\mathcal{G}^{\downarrow}_{A\cup S}$ and $\mathcal{G}^{\downarrow}_{B\,|\,S}$ are chain graphs. That is, $\mathcal{G}'$ is a chain graph.

Since S is a founder set of $\mathcal{G}'_{B\cup S} = \mathcal{G}^{\downarrow}_{B\,|\,S}$, it follows by Corollary 2 that $(\mathcal{G}'^{\downarrow}_{A\cup S})_S = \mathcal{G}'_S$. Therefore $\mathcal{G}'^{\downarrow}_{A\cup S} = \mathcal{G}^{\downarrow}_{A\cup S}$ implying that $\mathcal{G}'^{\downarrow}_{A\cup S}$ is an independence graph of $p_{A\cup S}$.

For $\mathcal{G}'^{\downarrow}_{B\cup S}$ to be an (not necessarily perfect) independence graph of $p_{B\cup S}$, we must require that

$$\left((\mathcal{G}^{\downarrow}_{B\cup S})_{\mathrm{An}(F\cup G\cup H)}\right)^m \subseteq \left((\mathcal{G}'^{\downarrow}_{B\cup S})_{\mathrm{An}(F\cup G\cup H)}\right)^m$$

for any $F, G, H \subseteq B$. Let $B' = \mathrm{An}(S)$ with respect to $\mathcal{G}^{\downarrow}_{B\cup S}$, $B'' = B \setminus B'$, and $N = \mathrm{ch}(S) \cap B'$, where $\mathrm{ch}(S)$ is determined with respect to $\mathcal{G}'^{\downarrow}_{B\cup S}$. Since $\mathcal{G}'^{\downarrow}_{A\cup S} = \mathcal{G}^{\downarrow}_{A\cup S}$ and by the construction of $\mathcal{G}^{\downarrow}_{B\,|\,S}$, it follows that $(\mathcal{G}'^{\downarrow}_{B\cup S})_{B''\cup S} = (\mathcal{G}^{\downarrow}_{B\cup S})_{B''\cup S}$. Further, since

$$(\mathcal{G}'^{\downarrow}_{B\cup S})_{B'\setminus S} = \left((\mathcal{G}^{\downarrow}_{B\cup S})_{B'\setminus S}\right)^m,$$

a difference between $((\mathcal{G}^{\downarrow}_{B\cup S})_{\mathrm{An}(F\cup G\cup H)})^m$ and $((\mathcal{G}'^{\downarrow}_{B\cup S})_{\mathrm{An}(F\cup G\cup H)})^m$ can only be caused by differences between the subgraphs induced by $S \cup N$, and, more specifically, only by differences in the two sets of edges between S and N. Let E' and E'' be these two sets pertaining to $\mathcal{G}^{\downarrow}_{B\cup S}$ and $\mathcal{G}'^{\downarrow}_{B\cup S}$, respectively. We note that $N \subseteq \mathrm{An}(S)$ in $\mathcal{G}^{\downarrow}_{B\cup S}$ and $N \subseteq \mathrm{ch}(S)$ in $\mathcal{G}'^{\downarrow}_{B\cup S}$. Thus, for $\mathcal{G}'^{\downarrow}_{B\cup S}$ to be an independence graph of $p_{B\cup S}$, we must require that $\alpha \sim \beta$ in $\mathcal{G}'^{\downarrow}_{B\cup S}$ whenever there are (not necessarily distinct) variables $\gamma, \delta \in N$ such that $\alpha \to \gamma$ and $\beta \to \delta$ and distinct variables $\alpha, \beta \in S$. However, by Corollary 3 it follows that $\alpha \sim \beta$ in $\mathcal{G}^{\downarrow}_{A\cup S}$, and (again), since $\mathcal{G}'^{\downarrow}_{A\cup S} = \mathcal{G}^{\downarrow}_{A\cup S}$, it follows that $(\mathcal{G}'^{\downarrow}_{B\cup S})_S = (\mathcal{G}^{\downarrow}_{B\cup S})_S$. Thus, the independence relations represented by $\mathcal{G}'^{\downarrow}_{B\cup S}$ are indifferent to the (lack of) direction of the edges between S and N. Therefore, $\mathcal{G}'^{\downarrow}_{B\cup S}$ and $\mathcal{G}^{\downarrow}_{B\cup S}$ are equivalent representations of the independence properties of $p_{B\cup S}$. Since the combination of $\mathcal{G}^{\downarrow}_{A\cup S}$ and $\mathcal{G}^{\downarrow}_{B\,|\,S}$ did not involve addition of edges, $\mathcal{G}'$ is perfect if both $\mathcal{G}^{\downarrow}_{A\cup S}$ and $\mathcal{G}^{\downarrow}_{B\,|\,S}$ are perfect. □

4.4 Some examples

In this section we present a few examples illustrating the modification of the independence graph when an edge is removed and show that the new decomposable cover may become significantly simpler.

The first example shows an independence graph consisting of a 4-cycle as illustrated in Figure 5. A decomposable cover may be generated by adding a fill edge between b and c. Each of the four edges of the independence graph is removable. Suppose we remove a—b. Using the above notation we get e.g. $A = \{b, c, d\}$, $B = \{a, c\}$ and $S = \{c\}$. Following the above results we should now determine a marginal independence graph and a conditional one, and then combine these into a new joint independence graph. This combination can involve three principally different sets of marginal and conditional graphs:

1. marginal graph for $A \cup S$ plus conditional one for $B\,|\,S$,
2. marginal graph for $B \cup S$ plus conditional one for $A\,|\,S$,
3. conditional graphs for $A\,|\,S$ and $B\,|\,S$ plus marginal one for S.

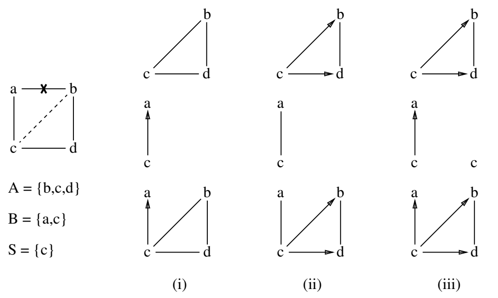


Figure 5: By removing $a — b$ the independence graph of the associated approximate joint belief can be created by combining the marginal independence graph for $A \cup S$ and the conditional one for $B\,|\,S$ (i), the marginal graph for $B \cup S$ and the conditional one for $A\,|\,S$ (ii), or the conditional graphs for $A\,|\,S$ and $B\,|\,S$ and the marginal one for S (iii).

Parts (i)–(iii) in Figure 5, respectively, show these three possibilities, where e.g. the upper graph in part (i) shows the marginal graph for $A \cup S$, the middle one the conditional graph for $B\,|\,S$, and the lower one the resulting combined graph. In this example the resulting new independence graphs appear to be equivalent, since, by Theorem 4, they are all equivalent to the graph obtained by substituting all the directed edges by undirected ones.

This small example shows an interesting feature, namely that the choice of decomposable cover affects the topology of the resulting independence graph. If, for example, the alternative (minimal) cover given by substituting $b \stackrel{\mathrm{f}}{—} c$ with $a \stackrel{\mathrm{f}}{—} d$ was applied, the resulting independence graph obtained by removing $a — b$ will be different, as shown in Figure 6. In both examples we observe that the fill edge becomes part of the independence graph.

A bit more complicated example is shown in Figure 7, where a minimal decomposable cover is given by addition of the fill edges indicated

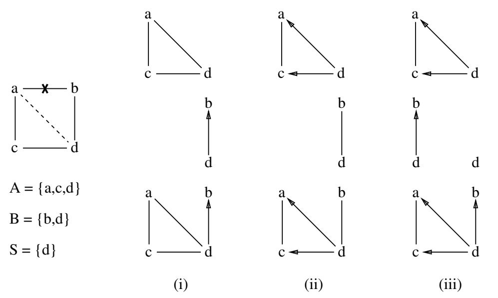


Figure 6: Similar to Figure 5 except that a different decomposable cover is used. This results in a different combined independence graph.

by the dashed lines. The three possible ways of generating the independence graph of the approximate joint belief are shown in parts (i)–(iii) corresponding to the possibilities 1–3 above (note: none of the combined graphs are minimally directed). In both of the conditional graphs for $A\,|\,S$ and $B\,|\,S$ one edge gets added by moralization. In the graph for $A\,|\,S$, $b — e$ gets added (see parts (ii)–(iii)), and in the graph for $B\,|\,S$, $d — e$ gets added (see parts (i) and (iii)). This means that the combined independence graph of part (iii) contains one edge more than those of parts (i)–(ii). Thus, one of the first two are preferable, since they are closer representations of the independence properties of the approximate joint belief. However, none of the graphs are perfect representations, since, by the marginal graphs for $A \cup S$ and $B \cup S$, $d \perp\!\!\!\perp e$ and $b \perp\!\!\!\perp e$, which all of the combined graphs fail to express. In fact, the three independence statements $d \perp\!\!\!\perp e$, $b \perp\!\!\!\perp e$, and $b \perp\!\!\!\perp d\,|\,(e, f)$ cannot simultaneously by represented in a single chain graph. (Note that the conditional independence of the variables between which the edge was removed given the remaining variables (the set called S above) of the relevant belief universe ($b \perp\!\!\!\perp d\,|\,(e, f)$ in this example) will always be properly represented in any of the resulting combined graphs.) The statement $b \perp\!\!\!\perp e$ is represented in the combined graph of part (i) while

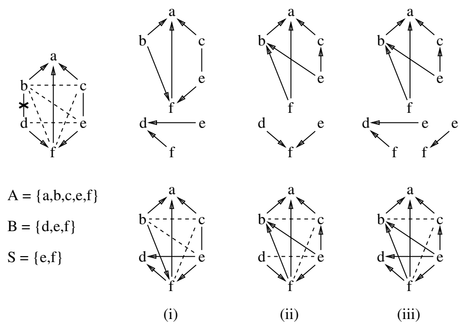


Figure 7: Removal of $b — d$ result in the corresponding independence graph of part (iii) if the two conditional graphs for $A \,|\, S$ and $B \,|\, S$ and the marginal one of S are combined, which contains one edge more than those of parts (i)–(ii). None of the graphs are perfect, since the fact that $d \perp\!\!\!\perp e$ and $b \perp\!\!\!\perp e$ are not represented.

$d \perp\!\!\!\perp e$ is not, and vice versa for part (ii).

In selecting one of the three alternatives for determining an independence graph of the approximate joint belief we should generally choose the conditional graph for which we have to add the fewest number of moral edges, since the more edges are added the less perfect the resulting combined independence graph will be. Thus alternative 3 in the above list will never provide better results than will alternatives 1 and 2. In all of the previous examples the decomposable cover of the approximate independence graph simply contains one edge less than the original cover, namely the edge removed. However, this is generally not true, since the removal of an edge may render some of the moral and fill edges superfluous. That is, a minimal cover of the approximate independence graph will not contain these edges. Figure 8 shows an example, where the fill edges $b \overset{\mathrm{f}}{—} c$ and $b \overset{\mathrm{f}}{—} f$ are rendered superfluous by the

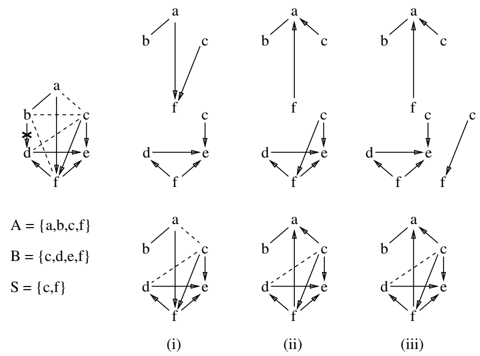


Figure 8: The independence graph obtained by combining the marginal graph for $A \cup S$ and the conditional one for $B \,|\, S$ is perfect (i). In all three cases the fill edges $b \overset{\mathrm{f}}{—} c$ and $b \overset{\mathrm{f}}{—} f$ are rendered superfluous.

removal of $b — d$. Again, for pedagogical reasons, minimal directedness has not been imposed. This example clearly demonstrates that alternative 1 (i.e., marginal graph for $A \cup S$ and conditional one for $B \,|\, S$) is preferable, since no moral edges get added in creating the conditional graph for $B \,|\, S$. A heuristic for choosing the optimal marginal graph upon which to perform conditioning with respect to S could be to select the one where $|\mathrm{pa}(S)|$ is smallest.

The final example in Figure 9 shows the resulting independence

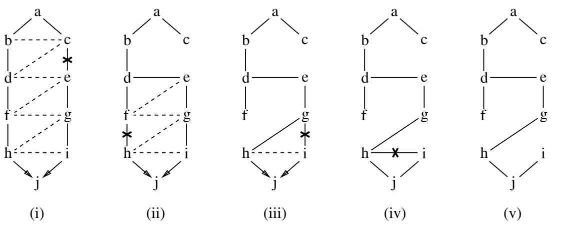


Figure 9: The removal of c—e renders $b \overset{\mathrm{f}}{—} c$ and $c \overset{\mathrm{f}}{—} d$ superfluous and makes $d \overset{\mathrm{f}}{—} e$ become part of the new independence graph. Similarly, the removal of f—h renders $e \overset{\mathrm{f}}{—} f$ and $f \overset{\mathrm{f}}{—} g$ superfluous and makes $g \overset{\mathrm{f}}{—} h$ become an edge of the new independence graph. By the removal of g—i, $h \overset{\mathrm{m}}{—} i$ becomes an edge of the new independence graph which in turn makes $h \to j$ and $i \to j$ undirected in a minimal directed representation.

graphs by a series of edge removals, where minimal directedness is imposed. The two first removals each render two fill edges superfluous. The third does not give rise to removal of superfluous edges, but it makes the moral edge between h and i become an edge of the new independence graph and hence, the directed edges $h \to j$ and $i \to j$ become undirected in a minimal directed representation which is also a decomposable cover of itself. Therefore the removal of any further edges cannot give rise to superfluous edges.

5 Selection criterion

Neglecting dependences (i.e., removing edges) has positive consequences in terms of cost reduction, but possible negative consequences in terms of erroneous inference. These two conflicting interests should be reconciled through a selection criterion trading off reduction of complexity for reduced 'information content' of the edge removed. To maximize the cost reduction we obviously want to remove edges belonging to belief universes with large state spaces, whereas to minimize the possible error, the dependences neglected should be the ones representing the weakest mutual information.

5.1 Information divergence

When removing an edge $\{\alpha, \beta\} \in E$ from a graph $\mathcal{G} = (V, E)$ as described above, we need a quantitative measure of the information lost by assuming $\alpha \perp\!\!\!\perp \beta \,|\, \gamma$, where $\gamma = V \setminus \{\alpha, \beta\}$. Put differently, we need an expression for the divergence between p and q, where p is the probability function for V before the removal of α—β and q is the probability function after the removal. In terms of p, q is given as

$$q = \frac{\sum_\alpha p \sum_\beta p}{\sum_{\alpha,\beta} p} = \frac{p(\beta,\gamma)p(\alpha,\gamma)}{p(\gamma)} = p(\alpha \,|\, \gamma)p(\beta \,|\, \gamma)p(\gamma), \tag{9}$$

implying that $\alpha \perp\!\!\!\perp \beta \,|\, \gamma$ under q.

Several divergence statistics exist. Perhaps the most well-known of these is the Kullback-Leibler information divergence statistic, also known as cross entropy or entropy distance. There are several good reasons for choosing this statistic as our measure of divergence, as we shall see below. The Kullback-Leibler divergence between two probability functions f and g is defined as

$$C(f,g) = \begin{cases} E\log(f/g) = \sum f \cdot \log(f/g) & \text{if } f > 0 \text{ and } g > 0 \\ 0 & \text{if } f = 0 \text{ and } g = 0 \\ \text{undefined} & \text{otherwise,} \end{cases} \tag{10}$$

and expresses the expected divergence between f and g, where the expectation is taken with respect to f. In the literature this measure is also frequently referred to as the directed divergence between f and g since, in general, $C(f,g) \neq C(g,f)$. Kullback (1959) refers to $C(f,g)+C(g,f)$ as the (undirected) information divergence. This measure, however, does not possess the as many properties as the undirected ones. When appropriate, the more precise notation $C(f,g;A)$ or $C(f_A,g_A)$ shall be used to indicate for which random variable (or vector) the divergence is taken.

In the special case where $f=p$ and $g=q$, with q given by (9), (10) is called the *conditional mutual information* between α and β given γ and is written

$$I(\alpha,\beta\,|\,\gamma) = C(p,q). \tag{11}$$

Properties of I are discussed in detail in Section 6.

The Kullback-Leibler divergence statistic has a number of important properties some of which shall be mentioned briefly in the following.

Theorem 8 *The Kullback-Leibler information divergence statistic satisfies*

(i) $C(f,g) \geq 0$,
(ii) $C(f_{A\cup B},g_{A\cup B}) = C(f_A,g_A) + C(f_B,g_B)$
if $A \perp\!\!\!\perp B$ *with respect to to both* f *and* g.

Statement (i) follows directly from Jensen's inequality and statement (ii) may be verified by simple algebraic manipulations.

The following important relationship between the maximum absolute divergence and the Kullback-Leibler divergence has been described by e.g. Csiszár (1967), Kemperman (1967), and Kullback (1967).

Theorem 9 *Let* V *be any set of discrete random variables, and let* f *and* g *be probability functions defined on* V. *Then for any subset* $A \subseteq V$

$$|f_A - g_A| \leq \sqrt{\tfrac{1}{2}C(f_V,g_V)}. \tag{12}$$

Proofs may be found in the above references, or in Whittaker (1989).

5.2 Reduction of complexity

From the above discussion it appears that when deciding which edge to remove we should consider not only the mutual information represented

by the edges and the immediate cost reduction (cf. (7)), but also the consequences of the removals in terms of rendering moral and fill edges superfluous (the removal of which further reduces the cost).

The total benefit in terms of cost reduction by removing an edge from a minimal decomposable cover, $\mathcal{C}$, is given by the cost difference between $\mathcal{C}$ and a minimal decomposable cover of the new moralized independence graph. Let the total cost reduction be

$$\Sigma(\alpha,\beta) = \sigma(\alpha,\beta) + \rho(\alpha,\beta), \tag{13}$$

where $\sigma(\alpha,\beta)$ is given by (7) and $\rho(\alpha,\beta)$ is the possible additional cost reduction obtained by rendering some (moral and fill) edges superfluous.

5.3 Combined measure

It might seem natural to choose an edge selection criterion based on a trade-off between the total cost reduction, $\Sigma(\alpha,\beta)$, obtained by removing α—β and the mutual information $I(\alpha,\beta\,|\,\cdot)$. However, an edge of a large belief universe might be rendered removable by first removing an edge yielding a relatively small cost reduction. Thus, given an amount by which the cost of a decomposable cover should be reduced, it is generally not possible to select an optimum sequence of edges to be removed using a selection scheme based on local considerations alone. To avoid a time-consuming search process one might wish to employ a heuristic selection criterion reflecting the trade-off between Σ and I. Such a heuristic might take the form

$$\lambda\Sigma - I, \tag{14}$$

where λ is a constant to be chosen such that $\lambda\Sigma$ is comparable with I. As $I \geq 0$ both terms in (14) are non-negative. Since we want I to be as small as possible and Σ as large as possible, the edge to be removed should be chosen such that the combined measure is maximized.

6 Propagation, accumulation, and attenuation of error

In this section we study the implications of edge removal in terms of propagation, accumulation, and attenuation of error. That is, we describe the circumstances under which an approximated belief potential does not influence the correctness of other belief potentials, prove that the overall error (information divergence) imposed by a series of edge removals equals the sum of the divergences associated with the individual edge removals, and prove that the influence of an approximation attenuates with increasing distance from the approximated belief universe.

6.1 The error may spread under arrival of posterior evidence

Now we investigate the effect of removing an edge $\alpha — \beta$ from a belief universe C_0 under subsequent absorption of evidence from a neighbouring universe C_i. In particular we show that if either C_i does not contain neither α nor β or at least one of α and β is contained in C_0 only, then the error induced by the removal of $\alpha — \beta$ does not spread to the neighbouring universes of C_0. In general, the error obviously spreads if this condition is not satisfied, since the joint potential for α and β is required for correct transferral of information from C_i via C_0 to any other neighbouring universe C_j when e.g. $\alpha \in C_i$ and $\beta \in C_j$.

Theorem 10 *Let ϕ be a joint belief for a junction tree $\Upsilon = (\mathcal{C}, \mathcal{S})$, where $C_0 \in \mathcal{C}$ is the unique universe containing $\{\alpha, \beta\}$, adj$(C_0) = \{C_1, \ldots, C_k\}$, and $S_j = C_0 \cap C_j$ for all $j = 1, \ldots, k$. Let further ψ be a joint belief for Υ given by (6), $\phi^*_{S_i}$ any potential on S_i $(1 \leq i \leq k)$ to be absorbed by C_0, and $\psi^*_{C_0}$, $\psi^*_{S_j}$, $\phi^*_{S_j}$, etc. the potentials for C_0 and S_j $(j = 1, \ldots, k)$ after the absorption. If either $\{\alpha, \beta\} \not\subseteq \bigcup_{j=1}^k S_j$ or $\{\alpha, \beta\} \cap S_i = \emptyset$, then $\phi^*_{S_j} \propto \psi^*_{S_j}$ for all $j = 1, \ldots, k$.*

Proof: By (2) we get

$$\psi^*_{S_j} = \sum_{C_0 \setminus S_j} \psi^*_{C_0} = \sum_{C_0 \setminus S_j} \psi_{C_0} \frac{\phi^*_{S_i}}{\psi_{S_i}} = \sum_{S_i \setminus S_j} \psi_{S_i \cup S_j} \frac{\phi^*_{S_i}}{\psi_{S_i}}.$$

By Lemma 3 we have that $\psi_{S_i} \propto \phi_{S_i}$, and since $\{\alpha, \beta\} \not\subseteq (S_i \cup S_j)$ or equivalently $(S_i \cup S_j) \subseteq C_0 \setminus \{\alpha\}$ or $(S_i \cup S_j) \subseteq C_0 \setminus \{\beta\}$, it follows similarly from Lemma 3 that $\psi_{S_i \cup S_j} \propto \phi_{S_i \cup S_j}$. Hence

$$\psi^*_{S_j} \propto \sum_{S_i \setminus S_j} \phi_{S_i \cup S_j} \frac{\phi^*_{S_i}}{\phi_{S_i}} = \sum_{C_0 \setminus S_j} \phi^*_{C_0} = \phi^*_{S_j}. \qquad \square$$

Note that in cases where external evidence on $A \subseteq C_0$ is entered into C_0, where $\{\alpha, \beta\} \cap A \neq \emptyset$, A can be perceived as the separator S_i from which C_0 absorbs, and hence the error spreads to the neighbours of C_0 only if the requirement $\{\alpha, \beta\} \not\subseteq \bigcup_{j=1}^k S_j$ or $\{\alpha, \beta\} \cap S_i = \emptyset$ is violated.

6.2 The Kullback-Leibler divergence can be computed locally

In this section we show that the Kullback-Leibler divergence $C(p_V, q_V)$, due to the removal of an edge $\alpha — \beta$ from a belief universe C_0, is equivalent to the marginal divergence $C(p_{C_0}, q_{C_0})$. Thus a vast reduction of the computational complexity can be achieved.

Theorem 11 *Let p_V be a consistent probability function for a junction tree $\Upsilon = (\mathcal{C}, \mathcal{S})$ with corresponding graph $\mathcal{G} = (V, E)$. Let $C_0 \in \mathcal{C}$ be the unique universe containing $\{\alpha, \beta\}$ and let q_V be a probability function for Υ given by (6) with proportionality constant 1. Then*

$$C(p_V, q_V) = C(p_{C_0}, q_{C_0}) = I(\alpha, \beta \mid \gamma),$$

where $\gamma = C_0 \setminus \{\alpha, \beta\}$ is the minimal set separating α from β in $\mathcal{G}' = (V, E \setminus \{\{\alpha, \beta\}\})$.

Proof: First note that q is consistent since p is consistent (cf. Corollary 1). With that in mind we easily get

$$\begin{aligned} C(p_V, q_V) &= \sum_V p_V \log \frac{p_V}{q_V} = \sum_V p_V \log \frac{\prod_{C\in\mathcal{C}} p_C/q_C}{\prod_{S\in\mathcal{S}} p_S/q_S} \\ &= \sum_V p_V \log \frac{p_{C_0}}{q_{C_0}} = \sum_{C_0} p_{C_0} \log \frac{p_{C_0}}{q_{C_0}} \\ &= C(p_{C_0}, q_{C_0}) = I(\alpha, \beta \,|\, \gamma). \end{aligned}$$

Now we prove that γ separates α from β in $\mathcal{G}'$ (i.e., the only path $\langle \alpha = \alpha_1, \ldots, \alpha_k = \beta \rangle$, $2 \le k \le |V|$, for which $S = \{\alpha_1, \ldots, \alpha_k\} \cap \gamma = \emptyset$ is $\langle \alpha, \beta \rangle$). Assume there is a path $\langle \alpha = \alpha_1, \ldots, \alpha_k = \beta \rangle$ between α and β such that $k > 2$ and $S \neq \emptyset$ (i.e., a path not including a vertex in γ). Since $\mathcal{G}$ is triangulated (i.e., all m-cycles for $m > 3$ have a chord), either $\{\beta, \alpha_2\} \in E$ or $\{\alpha, \alpha_{k-1}\} \in E$; otherwise there would be an m-cycle, $3 < m \le k$, without a chord. Thus either $\alpha_2 \in \gamma$ or $\alpha_{k-1} \in \gamma$ violating the assumption $S \neq \emptyset$.

Clearly γ is minimal as $\mathrm{adj}(\alpha) \cap \mathrm{adj}(\beta) = \gamma$ in both $\mathcal{G}$ and $\mathcal{G}'$. □

If interest in the belief potential p_{C_0} vanishes and no evidence is going to be received and retransmitted by C_0 then all edges which are unique to C_0 can be removed without affecting the belief potentials of other universes. This is a very useful result for belief updating in dynamic Bayesian networks (Kjærulff 1992a).

6.3 The Kullback-Leibler divergence is additive

When approximating a complex model by removing several edges, the following additivity property of Kullback-Leibler divergence is very important. Assume that k edges are removed sequentially from a junction tree Υ_0 with probability function p_0, and let Υ_i and p_i be, respectively, the junction tree and the probability function after the removal of the i^{th} edge. Then the Kullback-Leibler divergence between p_0 and p_k equals the sum over all $i = 1, \ldots, k$ of the divergence between p_{i-1} and p_i.

Theorem 12 *Let $\{\Upsilon_i\}_{i=1}^k$ be a collection of junction trees with belief universes $\{\mathcal{C}_i\}$, where $\mathcal{C}_i = \{C_i^1, \ldots, C_i^{m_i}\}$. Let C_{i-1}^* be the unique universe in $\mathcal{C}_{i-1}$ containing $\{\alpha_i^1, \alpha_i^2\}$, and for $j = 1, 2$ let*

$$A_i^j = \begin{cases} C_{i-1}^* \setminus \{\alpha_i^j\} & \text{if } \nexists C \in \mathcal{C}_{i-1} : C_{i-1}^* \setminus \{\alpha_i^j\} \subset C, \\ \emptyset & \text{otherwise.} \end{cases}$$

Further, for $i = 1, \ldots, k$ and $j = 1, \ldots, m_i$ let the belief universes $\{\mathcal{C}_i\}$ and the probability functions $\{p_i\}$ of $\{\Upsilon_i\}$ be given as

$$\mathcal{C}_i = (\mathcal{C}_{i-1} \setminus \{C_{i-1}^*\}) \cup \{A_i^1, A_i^2\},$$

$$p_i(C_i^j) = \begin{cases} \dfrac{\sum_{\alpha_i} p_{i-1}(C_{i-1}^*) \sum_{\beta_i} p_{i-1}(C_{i-1}^*)}{\sum_{\alpha_i,\beta_i} p_{i-1}(C_{i-1}^*)} & \text{if } C_{i-1}^j = C_{i-1}^*, \\ p_{i-1}(C_{i-1}^j) & \text{otherwise.} \end{cases}$$

Then $C(p_0, p_k) = \sum_{i=1}^k C(p_{i-1}, p_i)$.

Proof: First we rewrite $C(p_0, p_k)$ slightly

$$\begin{aligned} C(p_0, p_k) &= \sum_V p_0 \log \frac{p_0}{p_k} \\ &= \sum_V p_0 \log \left[\frac{p_0}{p_1} \frac{p_1}{p_2} \cdots \frac{p_{k-1}}{p_k} \right] \\ &= \sum_{i=1}^k \sum_V p_0 \log \frac{p_{i-1}}{p_i}. \end{aligned}$$

Since p_{i-1} and p_i differ only in clique C_{i-1}^* we have

$$\sum_V p_0 \log \frac{p_{i-1}}{p_i} = \sum_{C_{i-1}^*} p_0 \log \frac{p_{i-1}}{p_i},$$

and further, since the clique marginals are 'backward compatible', i.e.,

$$p_i(C_i^j) = p_{i-l}(C_i^j), \quad i = 1, \ldots, k, \quad l = 1, \ldots, i, \quad j = 1, \ldots, m_i,$$

we get

$$\sum_{C^*_{i-1}} p_0 \log \frac{p_{i-1}}{p_i} = \sum_{C^*_{i-1}} p_{i-1} \log \frac{p_{i-1}}{p_i} = C(p_{i-1}, p_i). \qquad \square$$

Among the divergence statistics described in the literature there are some which have the additivity property and some which do not. See e.g. the work of Read & Cressie (1988) for an overview of divergence statistics and their properties.

6.4 The error attenuates under increasing distance

Intuitively it seems reasonable that the effect of removing an edge from a belief universe, C, attenuates with increasing distance from C. Below we show this formally, but first we prove a lemma stating that the Kullback-Leibler divergence decreases under partitioning.

Lemma 5 *Consider a belief universe C with p and q being probability functions on $\mathrm{Sp}(C)$, and let $S \subset C$. Then*

$$C(p_S, q_S) \leq C(p_C, q_C).$$

Proof: Let $\|S\| = k$ and $\mathcal{S}_1, \ldots, \mathcal{S}_k$ be disjoint subsets of $\mathrm{Sp}(C)$ such that $\bigcup_{i=1}^k \mathcal{S}_i = \mathrm{Sp}(C)$, $s_i = \bigvee_{x \in \mathcal{S}_i} x \in \mathrm{Sp}(S)$, $p(s_i) = \sum_{\mathcal{S}_i} p$, and $q(s_i) = \sum_{\mathcal{S}_i} q$ for all $i = 1, \ldots, k$. Consider the partition of C into S as a consecutive series of collapses of $\mathcal{S}_1, \ldots, \mathcal{S}_k$ starting by $\mathcal{S}_1$. The first partition obtained is thus $\{s_1, \mathrm{Sp}(C) \setminus \mathcal{S}_1\}$ representing a smaller space, say $\mathrm{Sp}(C')$. Obviously, $C(p_{C'}, q_{C'})$ equals $C(p_C, q_C)$ minus the divergence between p and q in subspace $\mathcal{S}_1$ plus the divergence between $p(s_1)$ and $q(s_1)$, i.e.,

$$C(p_{C'}, q_{C'}) = C(p_C, q_C) - \sum_{\mathcal{S}_1} p \log \frac{p}{q} + p(s_1) \log \frac{p(s_1)}{q(s_1)}. \qquad (15)$$

The second partition $\{s_1, s_2, \mathrm{Sp}(C) \setminus \mathcal{S}_1 \setminus \mathcal{S}_2\}$ in this process is obtained by collapsing subspace $\mathcal{S}_2$ into s_2, and the divergence between p and q after this collapse is calculated by substituting $C(p_C, q_C)$ by $C(p_{C'}, q_{C'})$ and subscript 1 by subscript 2 on the right-hand side of (15). Continuing this process we find

$$\begin{aligned} C(p_S, q_S) &= C(p_C, q_C) - \sum_{i=1}^k \left[\sum_{\mathcal{S}_i} p \log \frac{p}{q} - p(s_i) \log \frac{p(s_i)}{q(s_i)} \right] \\ &= C(p_C, q_C) - \sum_{i=1}^k \left[p(s_i) \sum_{\mathcal{S}_i} p(x \,|\, s_i) \log \frac{p(x \,|\, s_i)}{q(x \,|\, s_i)} \right], \end{aligned}$$

where the inner-most sum is equal to the Kullback-Leibler divergence between $f_i(x) = p(x \,|\, s_i)$ and $g_i(x) = q(x \,|\, s_i)$ in space $\mathcal{S}_i$ since $\sum_{\mathcal{S}_i} f_i = \sum_{\mathcal{S}_i} g_i = 1$. Thus

$$C(p_S, q_S) = C(p_C, q_C) - \sum_{i=1}^k p(s_i) C(f_i, g_i) \leq C(p_C, q_C)$$

since both $p(s_i)$ and $C(f_i, g_i)$ are non-negative. $\square$

Based on that lemma we can now prove the following theorem.

Theorem 13 *Let $\Upsilon = (\mathcal{C}, \mathcal{S})$ be a consistent junction tree with probability function p, and let q be the consistent probability function for Υ obtained by substituting p_{C_0} with the probability function q_{C_0} and performing belief updating in Υ. Then for any path $\langle C_0, \ldots, C_k \rangle$ in Υ*

$$C(p_{C_0}, q_{C_0}) \geq C(p_{C_1}, q_{C_1}) \geq \cdots \geq C(p_{C_k}, q_{C_k}).$$

Proof: It suffices to consider the case $C(p_{C_1}, q_{C_1}) \leq C(p_{C_0}, q_{C_0})$. We have

$$C(p_{C_1}, q_{C_1}) = \sum_{C_1} p_{C_1} \log \frac{p_{C_1}}{q_{C_1}},$$

but since C_0 is the only universe which received evidence, then $q_{C_1} = p_{C_1} \dfrac{q_{S_1}}{p_{S_1}}$ and hence

$$C(p_{C_1}, q_{C_1}) = \sum_{C_1} p_{C_1} \log \frac{p_{S_1}}{q_{S_1}} = \sum_{S_1} p_{S_1} \log \frac{p_{S_1}}{q_{S_1}} = C(p_{S_1}, q_{S_1}).$$

By Lemma 5 we have that $C(p_{S_1}, q_{S_1}) \leq C(p_{C_0}, q_{C_0})$ which completes the proof. □

The principle of attenuation of evidence in a junction tree was pointed out by Hilden (1988).

7 Some examples

Since, from a theoretical point of view, not much can be said about the practical applicability of edge removal, we employ the method on a couple of real-world networks to gain some additional insight into the applicability of the method. Further, we compare the method with the annihilation technique (i.e., removal of probability mass) by Jensen & Andersen (1990).

7.1 Water treatment model

The first network is a probabilistic model of the biological processes of a water treatment plant (Jensen, Kjærulff, Olesen & Pedersen 1989). The model includes eight process parameters and is dynamic in the sense that, at any time instant, the model comprises a series of consecutive slices of time, where any two consecutive slices represent a time interval of 15 minutes. Each time slice t of the model describes the state of the dynamic process at time t and contains therefore eight nodes representing the random variables associated with the eight process parameters. Due to the relatively large time constants of the biological processes there are no within-slice edges, and thus, all probabilistic relationships between the process parameters appear as inter-slice edges (see Figure 10).

Due to the number of inter-slice edges and the ways they have been arranged, the complexity of the model become extremely high. After compilation in dHUGIN (Kjærulff 1992a) of a model including the first four time slices, the tables of the associated junction tree take up more than 43 megabytes of storage. The fact that the model has been considerably simplified to fit the capacity of workstation computers, indicates that a full-scale model probably would take up hundreds of megabytes of

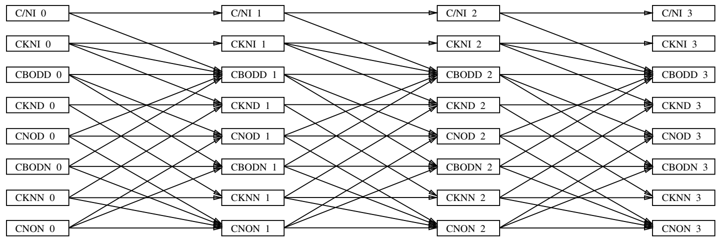


Figure 10: First four time slices of a dynamic Bayesian network model for the biological processes of a water treatment plant.

storage. This amount of storage is seldomly provided by today's workstation computers and is definitely not provided by today's low-cost personal computers which might be the target hardware for such a system. Therefore, further simplifications seem indispensable.

We have conducted two simplifications of the model (comprising the first four time slices) using the edge-removal technique. The first simplification (resulting in Model A) involves the removal of 43 edges selected according to a trade-off of the form (14), where we take the logarithm of the cost reduction, and the second simplification (resulting in Model C) involve 67 edges in addition (giving a total of 110 edges). The simplifications provided cost reductions by 90.2% and 96.5%, respectively. Approximations providing the same amounts of reductions using the method of Jensen & Andersen (1990) were also conducted for the sake of comparison. The resulting models are called Model B and Model D. The four models were evaluated by running them on three test cases discussed by Jensen et al. (1989) and comparing the resulting marginal distributions for the 32 variables with the corresponding correct distributions.

By recording the square error and the maximum absolute deviation for each variable, we can compute the average square error and the minimum, average and maximum of the maximum absolute deviation among the variables, and thereby get a summary of the merits of the individual models. Table 1 shows these figures for each model and each case, where Model A and Model B should be compared, and Model C and Model D should be compared.

Water treatment: all variables					
Model	Case	Min max dist	Ave max dist	Max max dist	Ave sq err
A	1	0.000000	0.000000	0.000000	0.000000
	2	0.000000	0.000000	0.000000	0.000000
	3	0.000000	0.000000	0.000000	0.000000
B	1	0.000000	0.000429	0.001371	0.000001
	2	0.000000	0.001889	0.007110	0.000014
	3	0.000000	0.359477	0.949389	0.469538
C	1	0.000000	0.000002	0.000061	0.000000
	2	0.000000	0.000000	0.000003	0.000000
	3	0.000000	0.006908	0.030908	0.000253
D	1	0.000000	0.003716	0.013097	0.000057
	2	0.000000	0.012682	0.040817	0.000595
	3	0.000000	0.360060	0.950088	0.474214

Table 1: Comparison of the edge-removal and the annihilation techniques for the water treatment model.

As indicated in Table 1 none of the posterior probabilities for the 32 process parameters differed more than $5 \cdot 10^{-7}$ from the 'correct' one when Model A was applied. When, however, the corresponding annihilation approximation was applied (Model B), Case 3 gave rise to extremely unreliable results. For Case 1 and Case 2, Model B provided quite acceptable result results. By the 96.5% reduction of cost, marginally significant errors can be registered for Case 3 under Model A. However, since the variables have at most 4 states this error can hardly be considered significant. The same is true for Case 1 and Case 2 under Model D, but for Case 3, Model D is useless.

The results provided by edge removal are not surprising since for the approximation producing Model A the (accumulated) Kullback-Leibler deviation is only 2.2×10^{-5} and for the approximation producing Model C it is 1.0×10^{-4}. In both cases the maximum absolute deviation (cf. (12)) is less than 0.01 (under lack of posterior evidence). Note that we observe a maximum absolute deviation of 0.03 under Model C (Case 3).

7.2 Winter wheat model

The second example is a preliminary model for the influence of temperature, precipitation, solar energy, and spraying against mildew on winter wheat in terms of leaf area index, severity of attack of mildew, production of dry matter, and expected yield. The model consists of four consecutive submodels (time slices) each representing a distinct phase of development (see Figure 11).

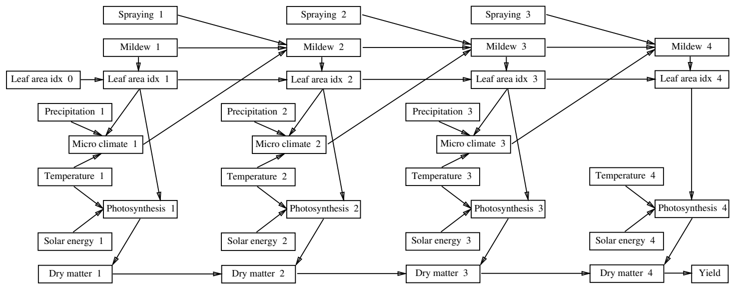


Figure 11: A Bayesian network model of growing of winter wheat.

We have conducted a single simplification of the model using the edge-removal technique. The resulting model, called Model A, was obtained by the removal of 26 edges. The simplification provided a cost reduction of 49.2%. An approximation providing the same amount of reduction using the annihilation technique was also conducted (this resulting model is referred to as Model B). These two models were evaluated by running them on three test cases. The resulting distributions for the four Mildew variables and the Yield variable were compared with the corresponding correct distributions. The reason for only considering these five variables is due to the fact that mildew is the only controllable parameter (via spraying) influencing the yield, and the reason for considering the yield should be obvious. The remaining variables determine the degree of attack of mildew and amount of dry matter produced (the accumulation of which determines the yield).

If we consider the approximated distributions of the Mildew variables, then we observe mostly insignificantly small deviations from the 'correct' distributions for both approximations. This pattern is broken in only one occasion, namely when running Model A with Case 1. Then the maximum absolute deviation for the distribution of variable Mildew_4 is about 0.1 (relative deviation 17.6%); see Table 2. The graphical display of the correct and approximated distributions in Figure 12 shows that even in this case the error is probably not serious since the shapes of the approximated and the correct distributions are roughly identical.

Winter wheat: Mildew						
Model	Case	Min max dist	Ave max dist	Max max dist	%	Ave sq err
A	1	0.006350	0.040844	0.103721	17.6	0.004556
	2	0.001520	0.006445	0.011031	13.7	0.000139
	3	0.000406	0.001486	0.002565	0.3	0.000005
B	1	0.006108	0.009248	0.012461	2.1	0.000151
	2	0.008700	0.015650	0.021074	3.4	0.000410
	3	0.000251	0.000399	0.000546	0.1	0.000000

Table 2: Comparison of the edge-removal and the annihilation techniques for the Mildew variables of the winter-wheat model.

A maximum absolute deviation of 0.012 for the Yield variable when applying Model A and Case 1 may seem to be insignificant (see Table 3). However, since the Yield variable has 86 states and since the distribution of the variable is fairly flat, the error is after all clearly significant. This is also indicated by the relative deviation of 41.8%, and it becomes very clear indeed when considering the graphical displays in Figure 13 of the correct and the approximate distributions of the variable. The approximate distribution (ii) of Figure 13 provided by Model A appears to be clearly bimodal and differs significantly from the correct one. The approximate distribution obtained by Model B appears to be a quite

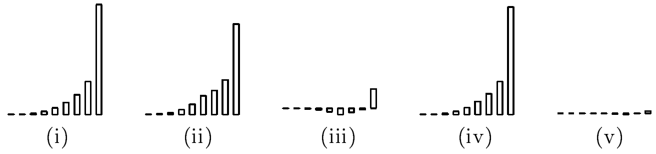


Figure 12: (i) The 'correct' distribution of the Mildew_4 variable under Case 1. (ii) The distribution under Model A. (iii) The absolute difference between (i) and (ii). (iv) The distribution under Model B. (v) The absolute difference between (i) and (iv).

close approximation of the correct one. Neither of the approximation techniques gave rise to serious deviations between the approximated distributions and the correct ones for the Yield variable when run with Case 2 and Case 3.

Winter wheat: Yield						
Model	Case	Min dist	Ave dist	Max dist	%	Sq err
A	1	0.000000	0.003537	0.012439	41.8	0.002616
	2	0.000000	0.001601	0.006361	23.0	0.000543
	3	0.000000	0.001345	0.004967	17.0	0.000377
B	1	0.000000	0.000566	0.001885	9.4	0.000053
	2	0.000000	0.000638	0.002676	5.6	0.000073
	3	0.000000	0.000404	0.002239	100	0.000038

Table 3: Comparison of the edge-removal and the annihilation techniques for the Yield variable of the winter-wheat model.

The rather discouraging results by the edge-removal technique in this example is not surprising since the accumulated Kullback-Leibler deviation for the 26 edges removed is about 0.2.

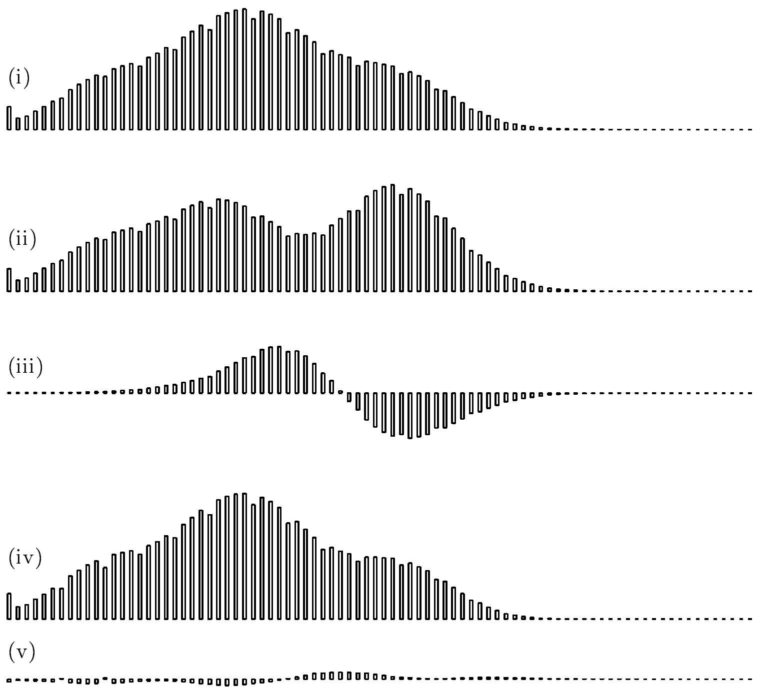


Figure 13: (i) The 'correct' distribution of the Yield variable under Case 1. (ii) The distribution under Model A. (iii) The absolute difference between (i) and (ii). (iv) The distribution under Model B. (v) The absolute difference between (i) and (iv).

8 Concluding remarks

The present paper has presented a scheme for approximating Bayesian networks through edge removals. A decomposable hypergraph cover (triangulated graph), of the network to be approximated, is first found and a corresponding junction tree is constructed along with an associated potential representation of the probability function. The removal of an edge amounts to enforcing additional conditional independence assumptions, expressed through modification of the potential representation, the junction tree, and the independence graph. We showed that although the initial independence graph is directed and acyclic, removal of edges from a corresponding triangulated graph may result in an independence graph being a chain graph. Further, using the modified independence graph, edges rendered superfluous by an edge removal are automatically identified. This feature is important, since the computational bottleneck in determining which edge to remove is calculation of the mutual information represented by candidate edges.

The Kullback-Leibler information divergence statistic was employed as the measure of mutual information. This measure has a number of important properties with respect to edge removal. We showed that the error induced by the removal of an edge affects only the belief universe (clique) from which the edge is removed, provided no evidence enters the universe (after the edge has been removed) from neither neighbouring universes nor the 'outside world'. Even in the presence of such posterior evidence, the error need not spread to neighbouring universes; however, it often does. As a consequence of the error being local, the Kullback-Leibler divergence can be computed using the (normalized) belief potential attached to the relevant universe instead of the (normalized) joint system belief. We also showed that the total divergence by a series of edge removals equals the sum of the divergences associated with the individual edge removals.

Since the key indicator associated with an approximation is most often the maximum absolute error, inequality (12) is essential, as it expresses the relationship between the Kullback-Leibler divergence and the maximum absolute error. However, the inequality only deals with the case where no posterior evidence has been received. Under arrival of posterior evidence, the inequality can only be used as a rough guideline (cf. the discussion at the end of Section 7.1).

An important feature of all approximation schemes performing local approximations (i.e., approximating single belief universes) is given by the fact that the effect of an approximation attenuates with increasing distance from the belief universe subjected to the approximation. This feature is especially important in connection with dynamic Bayesian networks (Kjærulff 1992a, Kjærulff 1993). An additional property of the presented scheme, is the property of errors remaining localized in absence of posterior evidence and, under certain conditions, even in presence of posterior evidence. That is, the belief potentials of the separators are unaltered by the removal of an edge. The approximation method by Jensen & Andersen (1990) does not have this property.

As was clearly demonstrated by the example in Section 7.1, the edge removal approach may in some cases be significantly better than the annihilation approach of Jensen & Andersen (1990). In other cases, however, a comparison between the two may turn out to the disadvantage of the edge removal approach (cf. Section 7.2). Intuitively, this seems quite reasonable, since a model including a lot of edges with a very low information content will be almost equivalent to a model which lacks these edges, but it might be quite different from a model obtained by uniformly removing a corresponding amount of probability mass over the belief universes. On the other hand, the edge removal approach is unsuited in cases where there are no edges with low information content. Thus, to approximate a given network using these two methods, approximation by edge removal should be tried first and when all edges with low information content have been removed, the annihilation technique should take over.

Among directions for future research, the following two points are important. Assessment of a good upper bound on the error upon arrival of posterior evidence is crucial. Such a bound is not provided in the present paper, but it is important to address this problem. As mentioned above, the network to be approximated is first transformed to a corresponding junction tree which may increase the complexity of representation

quite seriously. Thus, another area of interest for future research is selection of edges to be removed based solely on the original (conditional) probabilities of a Bayesian network. That is, edges are removed before transformation to a junction tree representation.

Acknowledgements

I am indebted to Steffen L. Lauritzen whose comments and support during the preparation of this paper were indispensable. I also wish to express my gratitude to the ODIN group for providing a very stimulating environment. The research has been funded partly by the Danish Research Councils through the PIFT programme.

References

Andreassen, S. A., Woldbye, M., Falck, B. & Andersen, S. K. (1987). MUNIN – A causal probabilistic network for interpretation of electromyographic findings, *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pp. 366–372.

Cooper, G. F. (1984). NESTOR: A computer based medical diagnostic aid that integrates causal and probabilistic knowledge, *Research Report HHP-84-48*, Knowledge Systems Laboratory, Medical Computer Science, Stanford University, California.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks, *Artificial Intelligence* **42**: 393–405.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations, *Studia Scientiarum Mathematicarum Hungarica* **2**: 299–318.

Darroch, J. N., Lauritzen, S. L. & Speed, T. P. (1980). Markov-fields and log-linear models for contingency tables, *The Annals of Statistics* **8**: 522–539.

Dirac, G. A. (1961). On rigid circuit graphs, *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg* **25**: 71–76.

Frydenberg, M. (1989). The chain graph Markov property, *Scandinavian Journal of Statistics* **17**: 333–353.

Geva, R. Y. & Paz, A. (1992). Representation of irrelevance relations by annotated graphs, *Research Report 603*, Technion — Israel Institute of Technology, Department of Computer Science, Haifa.

Hilden, J. (1988). Journal of the Royal Statistical Society, Series B, 50(2): 198–199 (Discussion of the paper by Lauritzen and Spiegelhalter).

Højsgaard, S. & Thiesson, B. (1992). BIFROST — Block recursive models Induced From Relevant knowledge, Observations, and Statistical Techniques, *Research report R-92-2010*, Department of Mathematics and Computer Science, Aalborg University, Denmark.

Howard, R. A. & Matheson, J. E. (1984). Influence diagrams, *Principles and Application of Decision Analysis*, Strategic Decisions Group, Menlo Park, California.

Jensen, F. & Andersen, S. K. (1990). Approximations in Bayesian belief universes for knowledge based systems, *Proceedings of the Sixth Workshop on Uncertainty in Artificial Intelligence*, Association for Uncertainty in Artificial Intelligence.

Jensen, F. V. (1988). Junction trees and decomposable hypergraphs, *Research report*, Judex Datasystemer A/S, Aalborg, Denmark.

Jensen, F. V., Kjærulff, U., Olesen, K. G. & Pedersen, J. (1989). Et forprojekt til et ekspertsystem for drift af spildevandsrensning (an expert system for control of waste water treatment — a pilot project), *Technical report*, Judex Datasystemer A/S, Aalborg, Denmark. In Danish.

Jensen, F. V., Lauritzen, S. L. & Olesen, K. G. (1990). Bayesian updating in causal probabilistic networks by local computations, *Computational Statistics Quarterly* **4**: 269–282.

Jensen, F. V., Olesen, K. G. & Andersen, S. K. (1990). An algebra of Bayesian belief universes for knowledge based systems, *Networks* **20**(5): 637–659.

Kemperman, J. H. B. (1967). On the optimum rate of transmitting information, *Probability and Information Theory* pp. 126–169. Lecture notes in mathematics, Springer-Verlag.

Kim, J. H. & Pearl, J. (1983). A computational model for causal and diagnostic reasoning in inference systems, *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp. 190–193.

Kjærulff, U. (1992a). A computational scheme for reasoning in dynamic probabilistic networks, *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*, Association for Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, Inc., pp. 121–129.

Kjærulff, U. (1992b). Optimal decomposition of probabilistic networks by simulated annealing, *Statistics and Computing* **2**: 7–17.

Kjærulff, U. (1993). A computational scheme for dynamic Bayesian networks, *Research Report R-93-2018*, Department of Mathematics and Computer Science, Aalborg University, Denmark.

Kullback, S. (1959). *Information Theory and Statistics*, John Wiley, New York.

Kullback, S. (1967). A lower bound for discrimination information in terms of variation, *IEEE Transactions on Information Theory* **IT-13**: 126–127.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N. & Leimer, H.-G. (1990). Independence properties of directed Markov fields, *Networks* **20**(5): 491–505. Special Issue on Influence Diagrams.

Lauritzen, S. L. & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems, *Journal of the Royal Statistical Society, Series B* **50**(2): 157–224.

Lauritzen, S. L. & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative, *The Annals of Statistics* **17**: 31–57.

Lauritzen, S. L., Speed, T. P. & Vijayan, K. (1984). Decomposable graphs and hypergraphs, *Journal of The Australian Mathematical Society, A* **36**: 12–29.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks, *Artificial Intelligence* **29**: 241–288.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Series in Representation and Reasoning, Morgan Kaufmann Publishers, Inc.

Read, T. R. C. & Cressie, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer Series in Statistics, Springer-Verlag, New York, Inc.

Rose, D. J. (1973). A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations, *in* R. C. Read (ed.), *Graph Theory and Computing*, Academic Press, New York, pp. 183–217.

Spiegelhalter, D. J. (1986). Probabilistic reasoning in predictive expert systems, *in* J. F. Lemmer & L. M. Kanal (eds), *Uncertainty in Artificial Intelligence*, North-Holland.

Wen, W. X. (1990). Optimal decomposition of belief networks, *Proceedings of the Sixth Workshop on Uncertainty in Artificial Intelligence*, Association for Uncertainty in Artificial Intelligence, pp. 245–256.

Wermuth, N. & Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables, *Biometrika* **70**(3): 537–552.

Whittaker, J. (1989). *Graphical Models in Applied Multivariate Statistics*, Wiley & Sons, Inc.