Efficient Information Gathering in Discrete Bayesian Networks

Marie Lilleborge

Dissertation presented for the degree of
Philosophiae Doctor (PhD)

Department of Mathematics
University of Oslo
July 2016

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1796

ISSN 1501-7710

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

To my mother and to my daughter

Preface

"On this path no effort is wasted, no gain is ever reversed." (Bhagavad Gita)

A PhD is such a large project -lengthy in time and dominating in everyday life- that life gets woven into the project just as much as the project is taking its place in life. I started as a PhD student in August 2012: Moved to a new city, bought my first apartment and met a boy. After a couple of months, I also got a "roommate" at NR in a fellow PhD student, the talented and friendly Martin Jullum. The first paper was submitted in October 2013, but kept haunting me for two more years. I started the task that had been scaring me the most, namely "finding something useful in the Junction Tree Algorithm in order to come up with some clever idea". My boyfriend's inspiring younger brother Ådne was undergoing chemotherapy at Rikshospitalet while I was sorting out the ideas for the "strongest possible messages" for MTP_2-distributions. Torgeir finished his Master's degree, we moved to California and I experienced a rough start of pregnancy. After months of guilt for slow progress, I refused to let my PhD experience further delays when I was hit by a car as a pedestrian in a foreign country five months pregnant. My first two papers were published during my maternity leave. I loved my work, but hated how I experienced it as a competitor to my daughter and my family.

I have had the pleasure of having three supervisors, each taking on very different roles. Ragnar Hauge is especially good at the so-called "green phase" of brainstorming, and I have enjoyed our many discussions of ideas based on mathematical theory yet unfamiliar to one or both of us. Jo Eidsvik is someone I look up to due to his academic skills as well as his friendliness and his balance of life. Jo taught me how to write scientific papers, and the importance of explicitly expressing the ideas you want the readers to take from a bunch of equations. Arne Bang Huseby at department of Mathematics, UiO, has been "our man" at UiO, an expert of the rules, procedures and forms. I also want to thank Arne for interesting and useful meetings and e-mail exchanges.

Thanks to my family and friends for participating in building a wonderful patchwork of memories. Torgeir and Edda, mamma Helena, pappa Jørn, Stine and Karina, you are my everything. I wish I could mention every friendship supporting me the last four years. Among the many I find Gireeja, Solveig, Johanne & Vidar, Mimi, Roger, Kristine, Heidi and fellow phd students in statistics at UiO including Martin and Tonje. I would also like to thank NR and the SAND department for providing a great environment for my phd work. I have felt well taken care of from the first month to the last; especially thanks to Petter Abrahamsen. Also many thanks to Solveig Hofvind and Sofie Sebuødegård at Kreftregisteret, for being co-authors of my fourth paper, and for encouraging words inspiring me to look forward to the next chapter.

Marie Lilleborge, Oslo, July 2016

List of papers

Paper I

LILLEBORGE, M., HAUGE, R. & EIDSVIK, J. (2016a). Information Gathering in Bayesian Networks Applied to Petroleum Prospecting. *Mathematical Geosciences* **48**, 233–257

Paper II

LILLEBORGE, M. & EIDSVIK, J. (2015). Efficient designs for Bayesian networks with sub-tree bounds. *Statistics and Computing* , To appear

Paper III

LILLEBORGE, M. (2016). Efficient optimization with Junction Tree bounds in discrete MTP2 distributions. Tech. rep., Norwegian Computing Center

Paper IV

LILLEBORGE, M., HOFVIND, S., SEBUØDEGÅRD, S. & HAUGE, R. (2016b). Using Bayesian Networks to optimize performance of the Norwegian Breast Cancer Screening Program - a modelling study. *Submitted for publication in Statistics in Medicine*

Contents

1 Motivation

The last 25 years, we have experienced both amazing improvements within the technology of transmitting, storing and retrieving data as well as huge advances in statistics. Data in general is now more accessible as sensoring of different environments and automatic data gathering is increasing in popularity. However, these types of large data sets often contain inconsistent data; they have different types of variables, and might have lots of missing data as well. This means that the interpretation of the data is crucial to gain useful information, as well as the question of how to best use the data or information at hand. This challenge has introduced a growing popularity of statistics but also an increased interest in black box approaches which tries to mimic the data without any evaluation of uncertainty or variability.

In many applications, however, data is still costly, not easily collected and/or not available in large quantities. For petroleum exploration in the North Sea, drilling an exploration well could cost $\$100$ million and is limited due to seasonal constraints. In medicine, a test is associated with both economical costs as well as inconveniences for the patient. In these scenarios, interpretation of how the result of different data gatherings will update our view of the situation will help guide which observations is more informative and how data should be collected.

Modelling uncertainty is key to better understanding, as knowledge is a combination of facts and logical implications together with the establishment of what is unknown. A model should incorporate both the uncertainty resulting from lack of knowledge and the variability in the situation modeled in order to be a proper representation of the phenomenon. New information updates the model and could possibly reduce the uncertainty. Reduction of uncertainty in the model then mimics increased knowledge and better understanding, while the variability will always remain. Probability is the mathematical language of uncertainty, and through probabilistic models we can reason about how updates and learning propagates between correlated variables.

During the last years, researchers have excelled in building complex models to describe reality, and invented computational methods for inference in these models. The Bayesian Network models are a result of mathematical research since the 1980s, and are among the key inventions from statistics the last 30 years. It was established as a field by Judea Pearl, and among the major well-known contributors we also find Finn Jensen and Steffen Lauritzen. BNs are now widely applied; -in medicine, defence, petroleum exploration, web-services, robotics, social networks and forensic science, to list a few. Pourret et al. (2008) presents twenty real-life case studies from different fields, together with discussions about strengths and limitations of the BN models for the specific applications and in general.

BNs can be used to find a *diagnosis* or an explanation for observations. By observing symptoms (evidence X_E), computation of conditional probabilities helps infer the most probable state of the variable causing them. Similarly, one can use BNs to learn how a variable X_i depends on earlier in time occurring variables X_E, and using the current state $X_E = x_E$ to make a *forecast* about the future state of X_i. BNs for classification are learning the connection between covariates and labels in a labeled dataset to predict labels on unlabeled data. BNs are also good tools for data mining tasks and for risk analysis.

BNs are attractive models for encoding qualitative and quantitative information. The BN modelling phase can incorporate several experts and different types of data in a consistent model. BNs are convenient for modeling complex dependencies between several random variables, and allow the construction of intuitive and modular probability statements at the local level. They can model different covariance patterns for different types of variables. In fact, these models can account for any correlation structure within the variables. As a graphical model, the BN is also a convenient tool to visualize the probabilistic dependencies on the model.

The very limitation of BNs is the computational complexity; both of building the model as well as probability updates in the built network. However, restricted to networks built from data by guidance of experts, the resulting size stays within the computational limits as experts naturally form models that are tractable even for the human brain to evaluate at least superficially. Enormous graphical models are also formed automatically by software at corporations like Google, Amazon and Netflix, where preprocessing allows for approximate calculations with great success.

Whether data is cheap or costly, easily accessible or hard to collect, obtaining information from data requires interpretation and clever reasoning. Further, different sources provide different data, which means the choice of future observations stochastically determines the information gained. The optimal information gain depends on which information is most useful for the current application. These considerations make information gathering a rich field for statistical research.

2 Bayesian Networks

Bayesian Networks (BNs) are used in several applications like medicine, forensic science, sensor validation, terrorism risk management, robotics as well as the oil industry. A BN is a directed graphical model, a way to specify a joint probability distribution of several random variables. It consists of a directed graph describing the conditional probabilistic dependencies between the variables, and a set of Local Probability Distributions (LPDs) which parametrizes the full joint distribution. Inference in BNs is known to be $\mathcal{NP}$-hard, see Cooper (1990).

A Bayesian Network can be learnt from data, specified by an expert or a combination of the two. Cowell et al. (2007) split the development of a BN into three phases. First, the relevant variables are specified. Second, the dependence structure between the variables is specified. This is referred to as the qualitative stage, where the relevance of one variable to another is considered. The third phase is to assign component probabilities, the numerical values required to build the full model. This last step is referred to as the qualitative stage.

In the qualitative building stage, the graph allows for intuitive modelling by a single expert, or similarly, allows several experts to have transparent discussions in order to agree on a common model. The graph can also be learnt from data, see chapter 7 in Jensen & Nielsen (2007) or chapter 3 in Højsgaard et al. (2012). For the qualitative building stage, the simplest idea is to estimate the required conditional probabilities directly as the corresponding frequencies in the data or use the maximum likelihood approach. In some applications, a few or all conditional probabilities are known to the expert, and can be directly specified. The conditional probabilities can also be estimated by a Bayesian approach, see chapter 9 in Cowell et al. (2007). Building the graph is not the focus of this thesis, and the reader is referred to the text books Jensen & Nielsen (2007), Cowell et al. (2007) and Højsgaard et al. (2012) for further reading.

2.1 Directed Acyclic Graphs

Directed graphs are commonly used without a proper definition, we present a definition from Cormen et al. (2009).

Definition 1. *A directed graph G is a pair (V, E), where V is a finite set and E is a binary relation on V. The set V is called the vertex set of G, and its elements are called nodes. The set E is called the edge set of G, and its elements are called edges.*

The elements in E are ordered pairs of nodes, and if $e = (i, j) \in E$, there is an edge e from node i to node j. We say that i is a *parent* of j and that j is a *child* of i. A *root* is a node with no parents, and a *leaf* is a node with no children. To illustrate a directed graph, each node is drawn as a circle and each edge (i, j) as an arrow from i to j; like in Figure 2.1.

A *walk* (from n_1 to n_N) is a sequence of nodes $n_1, \cdots, n_N$ such that $(n_i, n_{i+1}) \in E \; \forall i < N$,

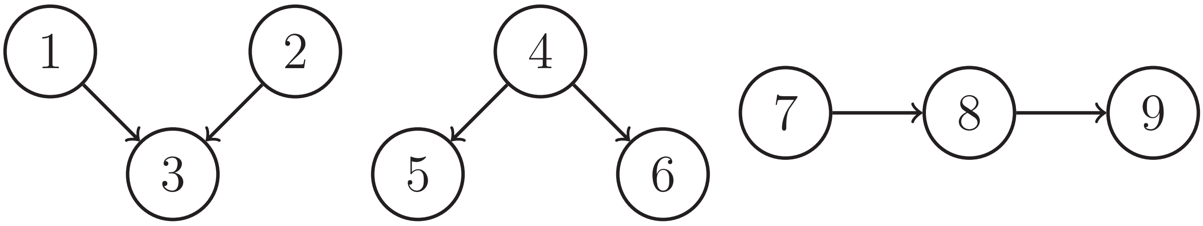

Figure 2.1: An example DAG $G = (V, E)$ with nine nodes $V = \{1, \cdots, 9\}$ and six edges $E = \{(1,3), (2,3), (4,5), (4,6), (7,8), (8,9)\}$. The DAG has three connected components: Collider $(\{1,2,3\}, \{(1,3), (2,3)\})$, Fork $(\{4,5,6\}, \{(4,5), (4,6)\})$ and Chain $(\{7,8,9\}, \{(7,8), (8,9)\})$.

and a *path* is a walk along distinct nodes.[1] If there is a path from node j to node k, we say that j is an *ancestor* of k, and k is an *descendant* of j. In Figure 2.1, we see that the leaf node 3 has two parents, $\text{Pa}(3) = \{1, 2\}$, and the same ancestors, $\text{Anc}(3) = \{1, 2\}$. Similarly, the root node 7 has one child, $\text{Ch}(7) = \{8\}$, and two descendants, $\text{Desc}(7) = \{8, 9\}$.

Let 2^V denote the power set of V, i.e. the collection $2^V = \{W : W \subseteq V\}$ of all subsets of V. The above family relations define functions from a node i to a set of nodes for which the family relation to i is met, namely

$$\begin{aligned}
\text{Pa} &: V \to 2^V \quad \text{s.t. } \text{Pa}(i) = \{j \in V \ : \ (j, i) \in E\}\,, \\
\text{Ch} &: V \to 2^V \quad \text{s.t. } \text{Ch}(i) = \{j \in V \ : \ (i, j) \in E\}\,, \\
\text{Anc} &: V \to 2^V \quad \text{s.t. } \text{Anc}(i) = \{j \in V : \exists \{k_\ell\}_{\ell=1}^m \text{ with } k_1 = j, k_m = i, (k_\ell, k_{\ell+1}) \in E \ \forall \ell\}, \\
\text{Desc} &: V \to 2^V \quad \text{s.t. } \text{Desc}(i) = \{j \in V : \exists \{k_\ell\}_{\ell=1}^m \text{ with } k_1 = i, k_m = j, (k_\ell, k_{\ell+1}) \in E \ \forall \ell\}.
\end{aligned}$$

It is common and practical to extend the definition of the functions to sets C by taking union over the evaluation for each element and excluding all variables already present in C, such that $\text{Anc}(C) = \left(\bigcup_{k \in C} \text{Anc}(k)\right) \setminus C$ and similarly for the other functions. In Figure 2.1, this means e.g. $\text{Pa}(\{3, 5\}) = \{1, 2, 4\}$ and $\text{Desc}(\{7, 8\}) = \{9\}$.

A *cycle* is a path with the modification that the first and last nodes are equal. Whenever a directed graph has no directed cycles, it is called a Directed Acyclic Graph (DAG). All DAGs have a topological ordering of the nodes, i.e. a bijective numbering of the nodes $\ell : V \to \{1, \cdots, |V|\}$ such that $\ell(j) < \ell(k)$ for any edge $e = (j, k) \in E$. The following straightforward topological sort algorithm for the nodes of a DAG is from Cowell et al. (2007):

- Initialize a copy of the graph: All vertices are unnumbered, and $i = 1$
- While there are vertices in the graph:
  - Give number i to a vertex with no parents and delete it from the graph
  - Update $i \leftarrow i + 1$

Cormen et al. (2009) proves that a depth-first search finishes the nodes in an opposite topological order. That is, another way to perform topological sort on a set of nodes in a DAG is to do a

[1] In the literature one will also find definitions that says a path is a sequence of edges (where each end node matches the next start node) or an alternating sequence of edges and nodes(where each edge is preceded by the start node and succeeded by its end node). However, these definitions have no practical differences implied.

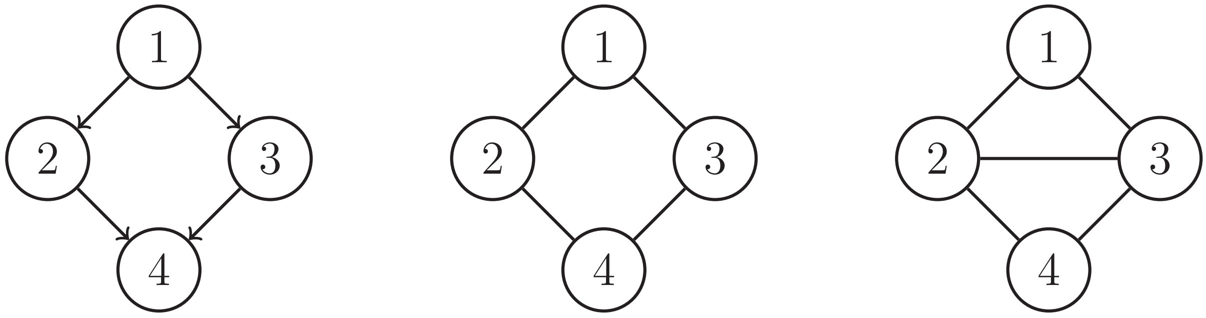

Figure 2.2: Left picture: An example DAG $G = (V, E)$ with $V = \{1, 2, 3, 4\}$, $E = \{(1, 2), (1, 3), (2, 4), (3, 4)\}$. Middle picture: Undirected version of the DAG. Right picture: Moral graph of the DAG.

depth-first search and insert each node at the top of a linked list whenever the search is finished processing it.

This thesis is about BNs, which are specified by directed graphs. However, the thesis is also about fast computation for BNs, and this will lead us to corresponding undirected graphs. An *undirected graph* has undirected edges $\{i, j\}$, which are usually visualized as lines between the corresponding pair of nodes. Some authors refer to undirected edges as links, perhaps to emphasize the difference of the relation introduced within the node-pair; namely the "equality" of the end points for an undirected edge compared to a directed one. An undirected edge can be interpreted as two directed edges, one in each direction, and be visualized as both directed edges. Similarly, a third option is to visualize the undirected edge as a double-headed arrow. In this thesis, we will visualize undirected edges as a line (with no arrowheads). If $\{i, j\} \in E$, we say that i and j are *neighbors*, and we define $\mathrm{Ne}(i) = \{j \in V \ : \ \{i, j\} \in E\}$ to be the set of neighbors of i. An undirected graph is *complete* if all pairs of nodes are neighbors. A set of nodes C in an undirected graph constitute a *clique* if all pairs of nodes in C are neighbors. If an undirected graph has a path between nodes i and j, we say that i and j are *connected*. The *undirected version* $(\tilde{V}, \tilde{E})$ of a directed graph (V, E) has the exact same nodes $\tilde{V} = V$ and each edge is represented without direction, $\tilde{E} = \{\{i, j\} : (i, j) \in E\}$.

A directed graph is also associated with another undirected graph called the moral graph, see Figure 2.2. To obtain the moral graph, a procedure called *moralization* is performed before the direction of edges are removed. For all triplets i, j, k where $j, k \in \mathrm{Pa}(i)$, an edge is added between j and k if there is not already one present (i.e. $(j, k) \in E$ or $(k, j) \in E$). This procedure ensures that all parents with a common child are married[2]. The *moral graph* is the undirected version of the graph after moralization.

According to Bondy & Murty (2008) a tree is an undirected graph which is connected and acyclic. It is easy to see that any pair of nodes in a tree are connected by exactly one path, and some references prefer this equivalent definition. As a simple example of a tree, we have the star graph $\{\{1, \cdots, n\}, \{\{1, j\} \ : \ j \geq 2\}\}$ of size n, where a single center node is connected to all other nodes and these other nodes are only connected to the center node. It is common to

[2] According to Oxford Dictionary, marriage is the "union of [two] partners in a relationship". In a graph, a relationship is to be joined together with an edge. For dynamic relationships, the reader is referred to Durrett (2007).

refer to a directed graph as a tree if it has a single root and its underlying undirected graph is a tree.[3]

A graph itself (directed or not) is connected if every pair of nodes are connected in the undirected version of the graph. A *subgraph* $(\tilde{V}, \tilde{E})$ of a graph (V, E) has a subset of the nodes $\tilde{V} \subset V$ and a restricted edge-set $\tilde{E} = \{(i, j) \in E \ : \ i, j \in \tilde{V}\}$. Any graph can be decomposed into *connected components* (subgraphs) where each node is represented in exactly one subgraph and all subgraphs are connected.

2.2 Bayesian Networks

In a *BN*, a DAG is used to express possible conditional independence assumptions among a set of random variables X_V. We let $X_A = [X_i]_{i \in A}$ denote the random vector indexed by an index set $A \subseteq V$, such that each entry X_i is a random variable for the index $i \in A$. In this thesis, we assume all random variables are discrete. Also, we let the assignment to a random variable be implicit, as we let $P\left(X_i | X_{\text{Pa}(i)}\right)$ denote $P\left(X_i = x_i | X_{\text{Pa}(i)} = x_{\text{Pa}(i)}\right)$ or $\mathbb{P}\left(X_V\right)$ denote $\mathbb{P}\left(X_V = x_V\right)$ for some implicit values of x_i and $x_{\text{Pa}(i)}$ or x_V. This is especially convenient when we are going to integrate out variables, i.e. sum over all possible assignments. The expected value $\mathbb{E}_{[X_V]} f(X_V)$ is explicitly written out as $\sum_{X_V = x_V} f(x_V) \mathbb{P}\left(X_V = x_V\right)$, but in the following it will be shortened down to $\sum_{X_V} f(X_V) \mathbb{P}\left(X_V\right)$.

The following definition is from Russell & Norvig (2003).

Definition 2. *A BN is a graph, consisting of a set of nodes $V = \{1, \cdots, n\}$ and a set of directed edges $E = \{e_i\}_{i=1}^{n_e}$ between pairs of the nodes. It is required that the graph has no directed cycles, i.e. it is a DAG. In addition, each node i represents a random variable X_i and has a set of LPDs $P\left(X_i | X_{Pa(i)}\right)$ associated with it. The full joint probability distribution over all the Random Variables represented in the network is*

$$\mathbb{P}(X_1, \cdots, X_n) = \prod_{i=1}^{n} P\left(X_i | X_{Pa(i)}\right). \tag{2.1}$$

Often in applications, one does not distinguish between the node i and the random variable X_i. For each node i and for each assignment to the random variables of its parents, $P\left(X_i | X_{\text{Pa}(i)}\right)$ is a probability distribution for the variable X_i, hence sums to 1. The *LPDs* are functions defining a local behavior (with respect to the parents) of a variable. We will see by conditioning and marginalization of the full distribution in (2) that actually the LPDs actually equals the corresponding conditional distributions, i.e. $\mathbb{P}\left(X_i | X_{\text{Pa}(i)}\right) = P\left(X_i | X_{\text{Pa}(i)}\right)$. We continue to refer to the LPDs as they are the defining pieces of the full distribution.

Observe by summing out variables in the opposite topological order that any set of nodes C has

$$\sum_{X_C} \prod_{k \in C} P\left(X_k | X_{\text{Pa}(k)}\right) = 1 \tag{2.2}$$

[3] Some references even allow all edges to point in the opposite direction (the somehow contradictory directed edges “towards the root”) and distinguish the two types of directed trees as in-trees (edges into root) and out-trees (edges out from root).

(for any assignment to $X_{\text{Pa}(C)}$), and

$$\mathbb{P}\left(X_C\right) = \sum_{X_{\text{Anc}(C)}} \prod_{k \in C \cup \text{Anc}(C)} P\left(X_k | X_{\text{Pa}(k)}\right) . \tag{2.3}$$

In fact, combining the above equation with Bayes' theorem proves

$$\mathbb{P}\left(X_j | X_{\text{Pa}(j)}\right) = \frac{\mathbb{P}\left(X_j, X_{\text{Pa}(j)}\right)}{\mathbb{P}\left(X_{\text{Pa}(j)}\right)} = P\left(X_j | X_{\text{Pa}(j)}\right) ,$$

that is, each LPD $P(X_j | X_{\text{Pa}(j)})$ equals the corresponding conditional probability distribution.

Recall that an edge in the DAG encodes a possible conditional dependence relationship between two variables in the BN, as the edges determine the variables each factor depends on in the formula of Definition 2. Whether a set of variables actually are conditionally dependent of each other, is determined by the parameters in the LPDs. Let $i \perp j$ denote if the graph ensures that two Random Variables X_i, X_j are independent ($\mathbb{P}\left(X_i\right) = \mathbb{P}\left(X_i | X_j\right)$), and $i \perp j \mid k$ if the graph ensures that X_i, X_j are conditionally independent given X_k ($\mathbb{P}\left(X_i | X_k\right) = \mathbb{P}\left(X_i | X_j, X_k\right)$). Correspondingly, we let $i \not\perp j$ denote that the graph does not encode that the Random Variables X_i, X_j are independent, as well as for $i \not\perp j \mid k$ in the conditional case.

The Bouncing Ball Algorithm in Jordan (t.a.) is an algorithm for finding all conditional independence relationships in a DAG. This algorithm is equivalent to the more tedious routine of using Bayes Rule on the general joint probability distribution as found in Definition 2 to check each possible independence statement in the given graph. Another algorithm for checking conditional independency statements is d-separation, and is based on a generalization of the three possible types of three-node interactions. We have seen the categorical three-node-interactions in Figure 2.1:

1. The "Collider" visualized by nodes $1, 2, 3$,
2. The "Fork" visualized by nodes $4, 5, 6$,
3. The "Chain" visualized by nodes $7, 8, 9$.

For any edge (i, j) we always have $i \not\perp j$, and for any two nodes k, l in different connected components we have $k \perp l$. For Figure 2.1, we also have $1 \perp 2$ and $1 \not\perp 2 \mid 3$ for the Collider, $5 \perp 6 \mid 4$ and $5 \not\perp 6$ for the Fork, $7 \perp 9 \mid 8$ and $7 \not\perp 9$ for the Chain.

Note that we let $\perp$ and $\not\perp$ denote independences implied by the graph, and in addition there are always LPDs that makes independences not ensured by the graph. In fact, if we let X_i or X_j be deterministic, the pair will be (conditionally and non-conditionally) independent. This however, does not prevent us from representing their joint distribution by a DAG where X_i and X_j are connected by an edge (X_i, X_j). The edge between X_i and X_j just allow for probabilistic dependence between them.

The following Theorem from Russell & Norvig (2003) describe the two standard conditional independence relations that are characteristic for BNs.

Theorem 1. *If the distribution function is positive ($\mathbb{P}\left(X_V\right) > 0$ for all assignments), then:*

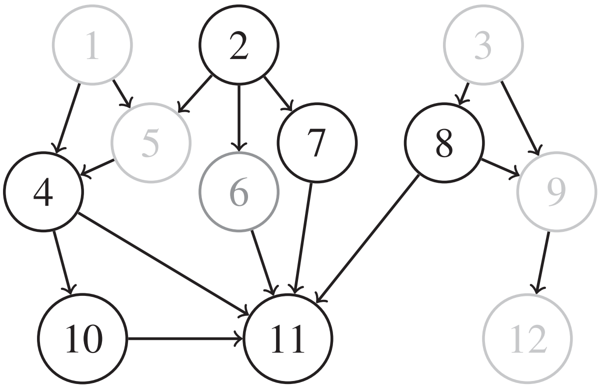

Figure 2.3: A BN with 12 nodes. The Markov Blanket of the dark gray node 6 is given black color, while the remaining nodes are in light grey. Theorem 1 gives two independence statements for node 6, namely $6 \perp 1, 5, 3, 9, 12 \mid 2, 4, 7, 8, 10, 11$ and $6 \perp 1, 3, 4, 5, 7, 8, 9, 10, 12 \mid 2$.

- *A Random Variable is conditionally independent of its non-descendants, given its parents.*
- *A Random Variable is conditionally independent of all other nodes in the network, given its Markov Blanket. The Markov Blanket of a node is the set consisting of its parents, its children and the parents of its children.*

The Markov Blanket of a node is illustrated in for an example DAG in Figure 2.3. The Theorem can be proved by applying Bayes rule on the joint probability distribution as found in Definition 2. For a further introduction to BNs, I recommend Jensen & Nielsen (2007), Cowell et al. (2007) or Koller & Friedman (2009) as textbooks purely on graphical models, or the broader Artificial Intelligence textbook Russell & Norvig (2003). I also recommend Bondy & Murty (2008) as a purely graph theoretic book[4] placing directed and undirected graphs in a more general framework.

[4] This book is free from probability distributions except for a chapter on random graphs, a concept out of scope for this thesis.

3 Junction Tree Algorithm

The Junction Tree Algorithm (JTA) is commonly considered the most efficient way to calculate a series of queries (probability statements) for a given graphical model, like a BN. The JTA was originally developed by Lauritzen & Spiegelhalter (1988), and has since then been established as the standard BN inference engine. There are several good JTA packages or open source implementations available. The JTA can be viewed as an improvement on the more intuitive Variable Elimination (VE) algorithm.

3.1 Variable elimination - a simpler inference engine

Assume we want to calculate the conditional distribution $\mathbb{P}\left(X_R|X_B\right)$ of the variables X_R for a given assignment to the variables in X_B. This instruction to calculate a given probability is referred to as a *query*. Note that potentially a query could have $B = \emptyset$, which instructs a marginalization from $\mathbb{P}\left(X_V\right)$ to $\mathbb{P}\left(X_R\right)$. On the other side, R could contain one or more nodes. Also, the JTA can return a representation of the conditional joint distribution of X_R or evaluate it for a given assignment $X_R = x_R$.

Recall from Chapter 2 that the full joint distribution of the variables in a BN is a product of factors $P(X_i|X_{\text{Pa}(i)})$, where each factor also is referred to as the LPD of the variable X_i. Thus, the variable X_i appears both in its own LPD as well as in the LPDs of its children. These LPDs are our initial *tables*, where for a given node i, the ”table” $P(X_i|X_{\text{Pa}(i)})$ has an entry for each possible assignment to $\{i\} \cup \text{Pa}(i)$ from which we can read off the corresponding probability $\mathbb{P}\left(X_i|X_{\text{Pa}(i)}\right)$. Let $\mathcal{D}$ denote the set of tables, such that initially $\mathbb{P}\left(X_V\right) = \prod_{D\in\mathcal{D}} D$, where we implicitly select the entry of each table which is consistent with the assignment to all variables X_V. The first step of the VE-algorithm is to incorporate the evidence to the tables by deleting all entries not consistent with the evidence. That is, for every node $b \in B$ and for every table D in which X_b appears, update D to the smaller table just containing the entries where the assignment to X_b is consistent with the evidence assignment $X_B = x_B$ in the query. It is now as if the nodes $b,\ b \in B$ does not appear in any table, since no table has entries depending on the assignment to X_b. This procedure is called instantiating the evidence.

Jensen & Nielsen (2007) further describe the routine of VE as:

- Repeat until only variables $X_i,\ i \in R$ appear in the tables in $\mathcal{D}$:
  - Select a variable $X_i,\ i \notin R$ appearing in some table $D \in \mathcal{D}$
  - Let $\mathcal{D}_i$ be the set of tables $D \in \mathcal{D}$ in which X_i appears
  - Remove all tables $D \in \mathcal{D}_i$ from $\mathcal{D}$
  - Calculate the product of all tables $D \in \mathcal{D}_i$

  - Marginalize X_i out of the new table
  - Place the resulting table in $\mathcal{D}$

- Normalize the resulting (product of) table(s) to obtain the distribution $\mathbb{P}\left(X_R|X_B\right)$.

Note that the VE-algorithm does not provide any guidance towards which order the variables X_i should be eliminated, i.e. marginalized out.

3.2 Efficiency

In each cycle of the VE-algorithm, a variable X_i is marginalized out of the set of tables $\mathcal{D}_i$. If all variables are binary, a table where n variables appear has size 2^n. Let n_i be the total number of variables appearing in the tables in the collection $\mathcal{D}_i$. Calculating the product of all tables in $\mathcal{D}_i$ means constructing and assigning values to a table with an entry for each assignment to the n_i variables, and hence has exponential time complexity in n_i. The marginalization of each variable X_i introduces such a new table, which is potentially constructed from previously constructed tables together with some original LPDs.

Assume a directed out-star with n binary variables, see the left picture of Figure 3.1 for $n = 6$.

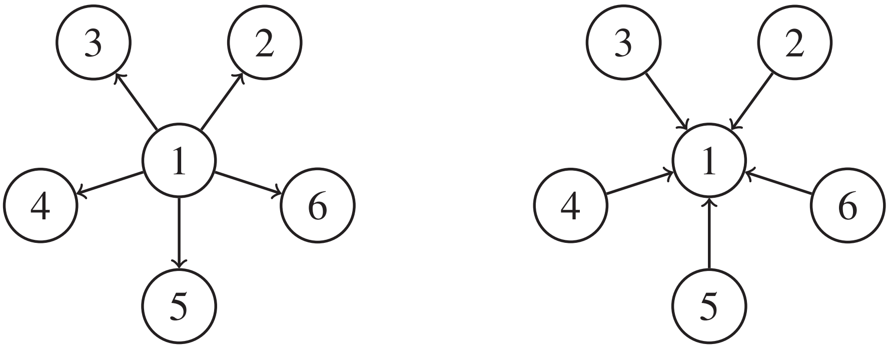

Figure 3.1: The directed star graphs of size 6. Left: Out-star. Right: In-star.

(Without loss of generality, we can assume this is the resulting graph after instantiating some evidence in some other variables.) Note that all LPDs depend on the center node X_1. We will first consider the elimination sequence $X_1, X_2, \cdots, X_n$. If X_1 is eliminated first, D_1 would contain all tables. The resulting table after X_1 is marginalized out would depend on all other variables, and have size 2^{n-1}. Before X_1 is marginalized out, each entry in the product-table is calculated as a product of one entry in each of the n LPDs. After X_1 is marginalized out, each entry in the resulting table is a sum of two of the product-table entries. We say that the time-complexity of constructing the resulting table after elimination of X_1 is $n \cdot 2^n + 2^n$, where the first term is for constructing the product-table, and the second term is for calculating the resulting table as 2^{n-1} sums of two entries. Each further marginalization will produce a table of half the size of the previous table until we are left with the desired (conditional) marginal probability as a table of size 1. Iteratively for $i = 2, \cdots n$, a table of size 2^{n-i} is calculated from the previous table of size 2^{n+1-i}, as each entry in the new table is a sum of two entries in the previous table. The iterative step i where X_i is eliminated has time complexity 2^{n+1-i}, and we end up with the total time-complexity $(n+2) \cdot 2^n - 2$ for the full VE-algorithm for the elimination sequence $X_1, X_2, \cdots, X_n$.

Consider the opposite order of the elimination sequence, namely $X_n, X_{n-1}, \cdots X_1$ for the out-star of size n. D_n would contain the LPD of X_1 and X_n, and the resulting table after X_n is marginalized out would depend only on X_1 and have size 2. The time-complexity of constructing this resulting table is $2 {\cdot} 2^2 {+} 2^2$, again with the first term for constructing the product-table and the second term for the pairwise sums. Correspondingly, each subsequent step $i = 2, \cdots n-1$ would have D_{n+1-i} containing the previous table and the LPD of X_{n+1-i}. The complexity of constructing the resulting table is again $2 \cdot 2^2 + 2^2$, and the resulting table would again only depend on X_1 after each step. Finally, step n is left with only the table constructed in step $n-1$ and the sum of the two entries is calculated with time complexity 2. The total time-complexity of the full VE-algorithm for the elimination sequence $X_n, X_{n-1}, \cdots X_1$ is $12n - 10$. We see that the out-star is an example where the elimination sequence has dramatic consequences on the time-complexity of the VE-algorithm. However, note that for some BNs, any elimination sequence leads to exponential time- and memory-complexity. As an example, the in-star of size n will have the first product-table contain all variables for any elimination sequence. The in-star is visualized in the right picture of Figure 3.1 for $n = 6$.

Both the sizes of the tables constructed by the VE-algorithm (complexity of memory needed) and the time-complexity of a full VE run depends on the elimination order in general. However, the VE does not provide any guidance for the order of the marginalizations or variable eliminations. In fact, computing the optimal variable elimination sequence is in general NP complete. If we are to compute several queries $\mathbb{P}\left(X_{A_j} | X_{B_j}\right)$ and can reuse the elimination sequence in some sense, it can obviously pay off spending some computational resources on finding a good elimination sequence. This is where the JTA comes in to play. It introduces an initial step where a computational object called a Junction Tree (JT) is constructed. The JT is an alternative representation of the joint distribution of the variables X_V, and it implicitly guides towards an elimination sequence. The problem of finding a good elimination sequence is now turned in to a problem of finding a good JT.

3.3 Standard JT construction

A JT is an undirected graph, more specifically a tree, and its nodes are representing a corresponding variable set. The JT we are going to construct will have nodes which represents a set of BN nodes. These sets will not be disjoint, but organize the BN nodes according to the probabilistic dependencies. In fact, each JT node will represent a table like in Chapter 3.1, and variables appearing in the same LPD of the BN will be appearing in a common JT node.

Almond & Kong (1991) present an alternative representation of the computational object which I myself prefer to the standard JT due to improvements in both calculation time and memory while not requiring significant changes to the JTA. One might argue that the difference is more or less an implementation detail, but in my perspective the theory in Almond & Kong (1991) simplifies not only the implementation but the theoretical presentation of the algorithm. However, since JTs as we will define in this section is the established standard, I find it most proper to start with this standard JT definition.

The construction of a JT for a given BN has an initial step where an undirected graph is constructed. This step is a combination of two general algorithms for graphs: First moralization of a DAG and then triangulation of the resulting undirected graph. The final step of JT construction includes a third general algorithm for graphs, namely finding a maximal weight spanning tree in an undirected graph. Moralization is known from Chapter 2 and efficient algorithms for finding a maximal weight spanning tree can be found in Cormen et al. (2009) (e.g. Kruskal's and Prim's algorithm in section 23.2). *Triangulation* is a procedure which adds edges until all cycles $v_{k_1}, \cdots, v_{k_\ell}$ of length ℓ in the undirected graph (V, L) does have a crossing edge $\{v_{k_i}, v_{k_j}\} \in L$ for a pair of indexes $1 \leq i,\ i+2 \leq j \leq n$. This crossing edge is commonly referred to as a *chord*, and a graph for which no edges are added during triangulation is called *chordal*. Note that all resulting graphs are chordal after triangulation. Optimal triangulation for JT construction is NP complete, originally proved in Yannakakis (1981). This is discussed in section 4.4.1 in Cowell et al. (2007), where also a one-step look ahead triangulation algorithm is presented. Triangulation is also discussed and attacked in section 4.6 of Jensen & Nielsen (2007).

Before introducing the full JT construction procedure, we will see how the first two steps (moralization and triangulation) illustrate the major difference of the similar-looking in-star and out-star from Figure 3.1. The moral graph of the out-star on the left side of Figure 3.1 is its undirected version, the undirected star on the left side of Figure 3.2. Correspondingly, the moral graph of the in-star on the right side of Figure 3.1 is the complete graph on the right side of Figure 3.2. This is related to the fact that there are efficient elimination sequences for the out-star but not for the in-star. Note that both undirected graphs in Figure 3.2 are triangulated, as the left graph has no cycles and the edge set of the right graph contains all possible chords for all of its cycles.

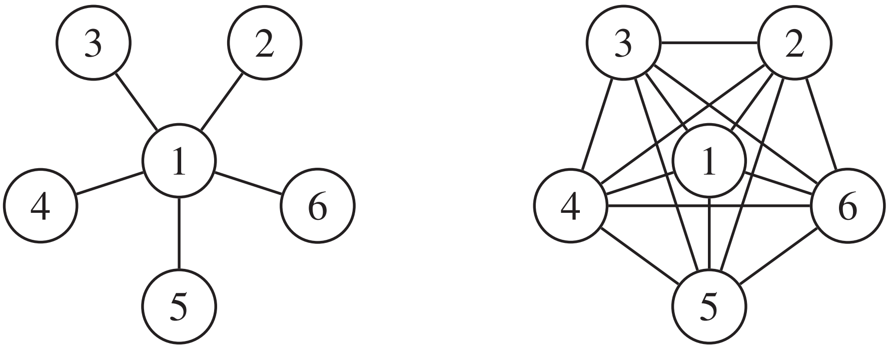

Figure 3.2: The undirected star graph of size 6 (left) and the complete undirected graph of size 6 (right).

JT is actually a general concept for undirected graphs, also used in relational databases[1]. Our definition is following Cowell et al. (2007).

Definition 3. *A JT is an undirected tree $(\mathcal{N}, L)$ whose nodes $N_i \in \mathcal{N}$ are associated with a variable set $\phi(N_i)$ each. A JT is required to have the running intersection property; for any two nodes N_i, N_j having a non-empty intersection $S = \phi(N_i) \cap \phi(N_j) \neq \emptyset$, this intersection S is*

[1] Join Trees is another name for Junction Trees

contained in the corresponding node set $\phi(N_k)$ for any node N_k on the (unique) path between the nodes N_i, N_j.

Note that some references, for example Jensen & Nielsen (2007), have definitions that require the variable set $\phi(N_i)$ to be the cliques of an underlying undirected graph.

In practice, the JT $\mathcal{T} = (\mathcal{N}, L)$ is constructed for a computational reason. The key point of the JT for efficient computation in BNs is that $\cup_{N_i \in \mathcal{N}} \phi(N_i) = V$ and that $\exists N_i \in \mathcal{N} : \{j\} \cup \text{Pa}(j) \subseteq \phi(N_i) \quad \forall j \in V$. This allows the JT to represent the joint probability distribution of all variables in the BN, and the latter requirement ensures that each LPD will have a JT node where all its variables are represented. There are in general several possible choices of a JT for a BN, for example the trivial $\mathcal{T} = (\{V\}, \emptyset)$. When the result of the moralization and triangulation is a complete graph, the trivial JT is the only option. For most BNs there are several possible JTs, see Figure 3.3. For efficient calculations, this choice matters. We will later see that it is desirable that each JT node represents a small subset of the BN nodes, and that the JT rather has a larger number of nodes. In fact for JTA, it is optimal to choose a JT where the set of nodes $\mathcal{N}$ represents (through ϕ) exactly the maximal cliques of a triangulated moralized version of the BN. In general, also this choice leaves several possibilities. We will see that a there is a link between a variable elimination sequence and a triangulation. However, finding the optimal JT is an NP-complete problem due to triangulation, so in applications one tries to find a ”good enough” solution. Note that in the following, for simplicity, we will not distinguish between the JT node N_i and its corresponding maximal clique $\phi(N_i)$ in the underlying BN. As is common in the literature, we will refer to the JT node as the node set N_i.

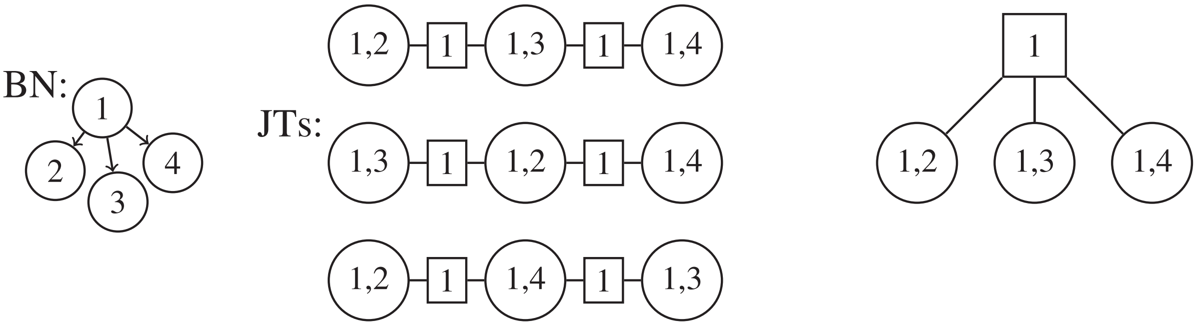

Figure 3.3: A BN with one parent and three children (left picture) has three possible standard JTs (middle pictures). Note that all three standard JTs have both edges with the same separator-set (node 1), and by merging equal separators as in Almond & Kong (1991) all three JTs result in the visually simpler AT on the right.

In the original formulation of the JTA, standard JTs as in Definition 3 are used. A standard JT is constructed from a BN by first moralizing the DAG, then triangulating the undirected version of the moralized graph, and finally presenting a maximal weight spanning tree from the complete graph whose nodes are the maximal cliques in the triangulated graph and where the weight of each edge (N_i, N_j) is $|N_i \cap N_j|$.

For the out-star (left side Figure 3.1), the triangulated moral graph is the undirected star (left side Figure 3.2) with maximal cliques $\{1, 2\}, \{1, 3\}, \cdots \{1, n\}$. Any pair of the maximal cliques have exactly one BN-node in common, which leads to a complete graph with node set $\{\{1, k\}_{k=2}^{n}\}$

and equal weights of all edges. Any spanning tree is therefore a maximal weight spanning tree, and we can choose the chain-graph with node set $\{\{1,k\}_{k=2}^{n}\}$ and edge set $\{\{\{1,k\},\{1,k+1\}\}\}_{k=2}^{n}$ as our JT. This is illustrated in Figure 3.3 for $n=4$. A more complex JT-construction process is visualized in Figure 3.4.

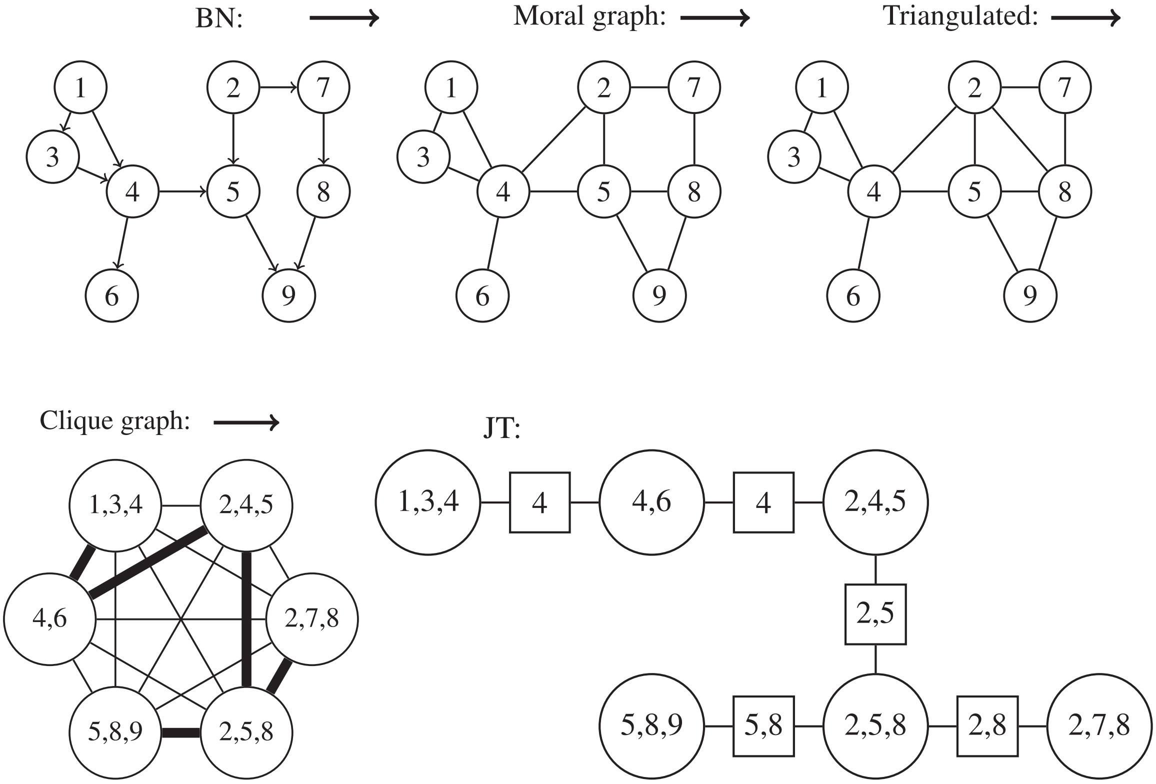

Figure 3.4: The process of constructing a standard JT (bottom right, separators visualized in a square on the corresponding edge) from a given BN (top left).

Finally each JT node N should be associated with a table D_N where the domain is all possible assignments to all variables X_N. Similarly as for VE, we want $\mathbb{P}(X_V) = \prod_{N\in\mathcal{N}} D_N$. As each LPD of the BN has a JT node where all its variables are represented, we select one such JT node for each LPD and let D_N be the product of the LPDs for which N was selected. Also let each separator (edge) $\{N_i, N_j\} \in L$ store a table D_{N_i,N_j} with domain all possible assignments to $N_i \cap N_j$. All entries of the separator tables can be set to have value 1 initially.

3.4 Message passing in a standard JT

Assume a JT $(\mathcal{N}, L)$ constructed from a BN (V, E) by moralization, triangulation and finally a construction of a maximal weight spanning tree in the complete clique graph. We refer to the elements of $\mathcal{N}$ as clique nodes N_i and the undirected edges $\{N_j, N_k\}$ of L as separators. We further assume that the clique nodes are numbered according to a topological ordering $N_1, \cdots, N_m$ and refer to N_1 as the root. Recall that for any node N_j, the (unique) path $(N_{k_1}, N_{k_2}, \cdots, N_{k_s}),\ k_1 = j, k_s = 1$ from N_j to N_1 has monotonically decreasing indexes $k_1 > k_2 > \cdots > k_s$. Define a neighbor-towards-root function $r : \mathcal{N} \to \mathcal{N} \cup \{\emptyset\}$ such that $r(N_j)$ is the unique lower-numbered neighbor of N_j for $j > 1$ and $r(N_1) = \emptyset$. Note that unless the JT is a chain, $\exists i \neq j$ such that $r(N_j) = r(N_i)$. That is, several JT nodes have the same neighbor-towards-root in general. We will treat $\emptyset$ as a fictious neighbor of the root N_1, catching the normalization constant as a result of the first sweep of the message passing. The message passing consists of two sweeps, first towards the root guided by the function r (from N_i to $N_j = r(N_i)$) and secondly away from the root as replies in opposite order of the towards root messaging (replies from N_j to all N_is such that $r(N_i) = N_j$).

Assuming a query $\mathbb{P}(X_A|X_B)$, the message passing in a JT is instructed as follows. Note that messages M are also tables.

- **Instantiate evidence:** For each variable $X_k,\ k \in B$, select a clique node N_j containing k. Set the entries of D_{N_j} to value 0 for all assignments to X_{N_j} violating the assignment to X_k in the evidence X_B. We continue working with $\prod_{j=1}^m D_{N_j}$ as an non-normalized representation of the joint conditional distribution for the non-evidenced variables. The normalization constant has the same value as the probability $\mathbb{P}(X_B)$ of the evidence, and will be calculated as a result of the first sweep of the message passing.

- **Message passing towards root:** For $j = m, \cdots, 1$, send message from node N_j: Collect the current table D_{N_j} and the incoming messages $M_{N_k}^{\rightarrow}$ in a temporary table ϕ_j, from which the outgoing message $M_{N_j}^{\rightarrow}$ to $r(N_j)$ is computed as a marginal of the variables $N_j \cap r(N_j)$ represented in both ends of the separator $\{N_j, r(N_j)\}$, namely

$$M_{N_j}^{\rightarrow} = \sum_{X_{N_j \setminus r(N_j)}} \phi_j, \qquad \phi_j = D_{N_j} \cdot \prod_{N_k \in \mathrm{Ne}(N_j) \setminus \{r(N_j)\}} \frac{M_{N_k}^{\rightarrow}}{D_{N_j,N_k}}.$$

  After the message $M_{N_j}^{\rightarrow}$ is sent (making $M_{N_j}^{\rightarrow}$ an incoming message to $r(N_j)$), reset the current table to $D_{N_j} = \frac{\phi_j}{M_{N_j}^{\rightarrow}}$.

- **Intermediate result:** After message passing towards root and the subsequent message $M_1^{\rightarrow}$ passed from the root N_1 to its fictious neighbor $\emptyset$, the probability of the evidence is collected as $\mathbb{P}(X_B) = M_1^{\rightarrow}$ and we have a normalized representation of the conditional joint distribution

$$\mathbb{P}\left(X_{V \setminus B}|X_B\right) = \prod_{j=1}^{m} D_{N_j} = \prod_{j=1}^{m} \mathbb{P}\left(X_{N_j \setminus r(N_j)}|X_{N_j \cap r(N_j)}\right).$$

- **Message replying outwards from root:** For $j = 1, \cdots, m$, send a reply from node N_j to each node N_k with $r(N_k) = N_j$, i.e. each node from which N_j received a message in the message passing towards root. If $j > 1$, reset D_{N_j} to ϕ_j calculated as a product of the incoming reply and the current table, i.e. $\phi_j = M_{r(N_j)}^{N_j \leftarrow} \cdot D_{N_j}$. After the update, the current table stores the conditional joint of its BN variables. Then, for each k such that $r(N_k) = N_j$, send as a reply the marginal of $X_{N_k \cap N_j}$

$$M_{N_j}^{N_k \leftarrow} = \sum_{X_{N_j \setminus N_k}} D_{N_j}$$

from N_j to N_k over the separator $\{N_j, N_k\}$. Let the separator store a copy $D_{N_i,N_j} = M_{N_j}^{N_k\leftarrow}$.

- **Result:** In the intermediate result, we used that the product is constant as long as one term is multiplied with the same amount as another term is divided by. We will use the same principle here, but as the tables of the clique nodes were not divided by the messages they send outwards from root, these messages were stored in the separators. For any sub-tree $\mathcal{T}_J = (N_J, L_J)$ where J is a collection of indexes such that $N_J = \{N_j \in \mathcal{N} \ : \ j \in J\}$ and $L_J = L \cap (N_J \times N_J)$, we have

$$\mathbb{P}\left(\cup_{j\in J} N_j \setminus B | B\right) = \frac{\prod_{j\in J} D_{N_j}}{\prod_{(N_j,N_k)\in L_J} D_{N_j,N_k}}.$$

  The simplest examples are single cliques $J = \{j\}$, or all variables $J = \{1, \cdots, m\}$. We will later use that this formula is correct both globally and locally.

The message passing is illustrated for two JT nodes in Figure 3.5.

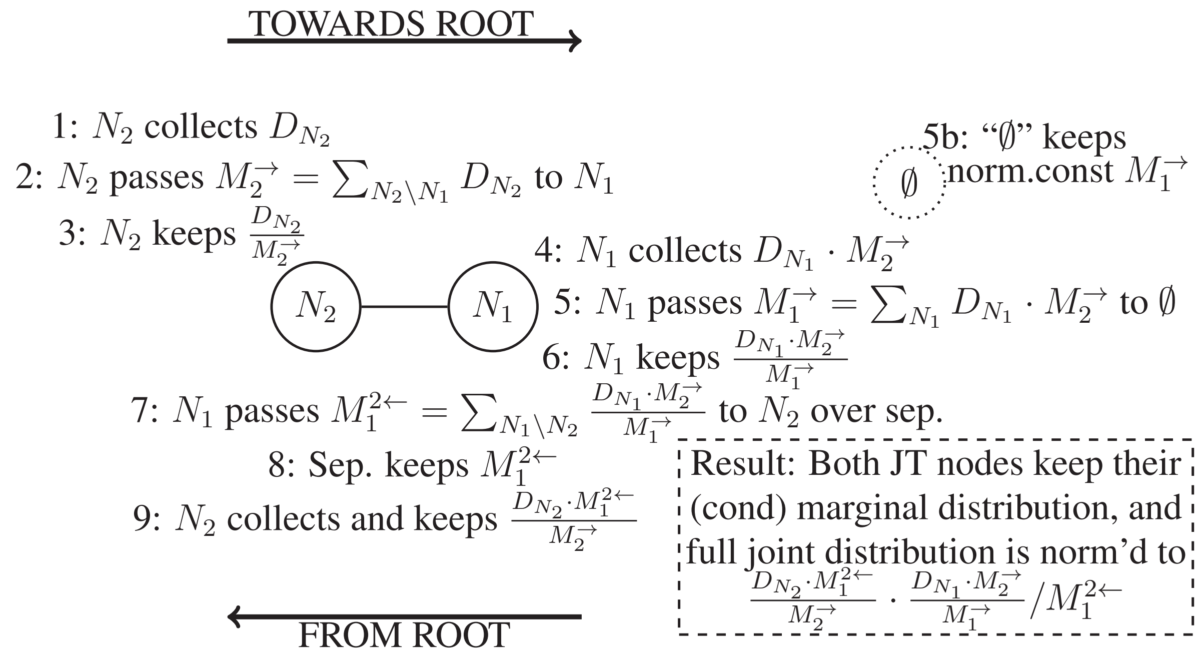
TOWARDS ROOT
1: N_2 collects D_{N_2}
2: N_2 passes $M_2^{\rightarrow} = \sum_{N_2 \setminus N_1} D_{N_2}$ to N_1
3: N_2 keeps $\frac{D_{N_2}}{M_2^{\rightarrow}}$
5b: "$\emptyset$" keeps norm.const $M_1^{\rightarrow}$
$\emptyset$
4: N_1 collects $D_{N_1} \cdot M_2^{\rightarrow}$
N_2
N_1
5: N_1 passes $M_1^{\rightarrow} = \sum_{N_1} D_{N_1} \cdot M_2^{\rightarrow}$ to $\emptyset$
6: N_1 keeps $\frac{D_{N_1} \cdot M_2^{\rightarrow}}{M_1^{\rightarrow}}$
7: N_1 passes $M_1^{2\leftarrow} = \sum_{N_1 \setminus N_2} \frac{D_{N_1} \cdot M_2^{\rightarrow}}{M_1^{\rightarrow}}$ to N_2 over sep.
8: Sep. keeps $M_1^{2\leftarrow}$
9: N_2 collects and keeps $\frac{D_{N_2} \cdot M_1^{2\leftarrow}}{M_2^{\rightarrow}}$
Result: Both JT nodes keep their (cond) marginal distribution, and full joint distribution is norm'd to
$\frac{D_{N_2} \cdot M_1^{2\leftarrow}}{M_2^{\rightarrow}} \cdot \frac{D_{N_1} \cdot M_2^{\rightarrow}}{M_1^{\rightarrow}} / M_1^{2\leftarrow}$
FROM ROOT

Figure 3.5: Message passing in a standard JT with two clique nodes. Note that in this picture, D_{N_1} and D_{N_2} refer to the tables untouched by message passing to bring intuition of how the potentials get distributed after a message passing routine.

To compare with the VE-algorithm for the out-star with n binary variables (left side Figure 3.1) and the corresponding chain-JT, each of the $2n$ collect-steps of the JTA has time-complexity $2 \cdot 2^2$ with a further marginalization of time-complexity 2^2. Further, $n-1$ separators stores a table of size 2, each introducing a time-complexity term 2. We end up with a total time-complexity of $26n - 2$, again linear in n and now presenting (conditional) marginals for all cliques and separators. Recall that if the (conditional) success probability of X_3 is a desired quantity, it is calculated as the sum of the two entries in $D_{\{1,3\}}$ where $X_3 = 1$ holds (namely $X_1 = 1, X_3 = 1$ and $X_1 = 0, X_3 = 1$).

In the following we will first discuss the similarities and differences between the two approaches, and then we will continue assuming the computational object as in Almond & Kong (1991) and refer to it as an Almond Tree[2] (AT) to avoid confusion.

3.5 AT construction

This paragraph presents an AT as a version of the standard JTs. Recall that a JT is an undirected tree where each of its nodes represents a clique N_i in an undirected graph constructed from the BN. Each of the edges (N_i, N_j) in the JT is associated with what we call a separator $S_{i,j}$ which represents the node set $S_{i,j} = N_i \cap N_j$. Sometimes several separators represent the exact same node set. In these cases it would be more efficient to run JTA on the corresponding AT. The out-star is again an obvious example where all separators are equal.

An AT is a computational object like a JT, except the separators are also viewed as (a special kind of) nodes which we will refer to as almond nodes. Assume we are given a JT and want to construct the corresponding AT. Then, we first expand each one-edge long path $N_i, (N_i, N_j), N_j$ between two neighboring cliques N_i, N_j in the original JT to a two-edge long path $N_i, (N_i, S_{i,j}), S_{i,j}, (S_{i,j}, N_j), N_j$, where $S_{i,j}$ is the separator $N_i \cap N_j$ associated with the edge (N_i, N_j). Consequently, equal separators are merged into a single separator in the AT. This kind of separator is called an almond (separator), and has a corresponding multiplicity which equals the original number of merged separators, or equivalently, one less than the number of its neighbors. When the equal separators $S_{i,j}$ are merged to one single separator S, all edges $(N_i, S_{i,j})$ are translated to (N_i, S) and duplicate edges are deleted. See Figure 3.3 for corresponding AT and standard JTs.

One can construct cases where one need to rearrange JT nodes on a path to avoid cycles when separators are merged, but this always corresponds to another choice of maximum weight spanning tree in the construction of the JT. This can be avoided by choosing maximum weight edges (separators) such that equal separators connect, as there is always such an maximal weight spanning tree alternative.

The improvement in memory usage for ATs compared to standard JTs is that there are fewer or the same amount of tables. A standard JT has one table for each clique node and one for each edge (separator). Similarly, an AT has one table for each clique node and one for each almond node (separator). As long as the standard JT has no edges with equal separator set, the set of tables are exactly the same for the corresponding AT and standard JT. In this case, the two structures have the same memory usage except for storing the tree-structure and the multiplicity of the almonds (negligible difference). However, in the case of at least one pair of equal separator set for edges, merging equal separators for the AT also means a reduction in the number of tables. Tables which otherwise would be equivalent to other tables are not constructed. This means less memory used and less computing time for message passing as there are fewer tables to update.

According to Almond & Kong (1991):

[2] According to Russel Almond's homepage, Finn Jensen also has also been referring to these trees as Almond Trees.

Definition 4. *Let $\mathcal{T} = (\mathcal{N}, L)$ be an undirected tree in which the nodes in $\mathcal{N}$ are labelled subsets of some index set V. The tree $\mathcal{T}$ is a Markov Tree if for any two nodes $N_1, N_2 \in \mathcal{N}$, any other node N_3 which lies on the path between them must satisfy $N_1 \cap N_2 \subseteq N_3$. An AT is a Markov Tree with the additional property that for every pair of neighboring nodes, one is a subset of the other.*

As pointed out in Almond & Kong (1991), the standard JTs with separators considered as nodes are a special case of ATs, since JTs are Markov (Spanning) Trees of the complete clique graph. In the following, we adhere to a distinction between almond nodes (separators in the AT) and clique nodes. We will specify the AT as $\mathcal{T} = (\mathcal{C} \cup \mathcal{A}, L)$ where the node set $\mathcal{N}$ is split into a disjoint union of clique nodes $\mathcal{C}$ and almond nodes $\mathcal{A}$. Correspondingly, almond nodes will be marked as squares and clique nodes as circles in our visualizations.

The above procedure describing how we can go from standard JT to AT is only presented to build understanding about the similarities of the objects. In practice, the AT is constructed from the BN in a procedure where the similarities to the VE-algorithm become clearer, since it relies on a variable elimination order. The AT construction algorithm relies on a choice of variable elimination order through a heuristic, as finding the optimal order is NP complete in general.

Almond & Kong (1991) provide the following argument for simple variable elimination order heuristics: Various variations of one step ahead algorithms “work optimal, or near optimal, in a large number of cases $(\cdots)$[and] takes less time. The fewest fill-ins heuristic is often as effective as the compound heuristics”. The fewest fill-ins heuristic iteratively from the current working-copy of the graph $(\mathcal{N}, L)$ selects node n with smallest fill-in number $|\,\{\{\ell, m\} \notin L \ : \ \{\ell, n\}, \{m, n\} \in L\}\,|$ as the next node to eliminate. We also chose to follow the fewest fill-ins heuristic in our JTA implementation for Lilleborge & Eidsvik (2015) and for the JTA-based calculations implemented for Lilleborge (2016).

The following AT construction algorithm is from Almond & Kong (1991) and starts with the DAG of the BN:

- Moralize the DAG
- Remove the direction of all edges and obtain an undirected graph $(\mathcal{N}^0, L^0)$
- Select a variable elimination order $j_1, \cdots j_{|V|}$ according to your favourite variable elimination order heuristic (e.g. fewest fill-ins). At step $k = 1, \cdots, |V|$:
  - The heuristic points to $j_k = n$ for a node n in the undirected graph $(\mathcal{N}^{k-1}, L^{k-1})$
  - Define $D_k = \{i \in \mathcal{N}^{k-1} \ : \ \{i, n\} \in L^{k-1}\}$ as the set of neighbors of node n in the current graph
  - Update graph by removing node n: $\mathcal{N}_k = \mathcal{N}_{k-1} \setminus \{n\}$, $L_k = \{\{i, m\} \in L_{k-1} \ : \ i, m \neq n\}$
- Build the full AT $\mathcal{T} = (\mathcal{C} \cup \mathcal{A}, L)$ by moving backwards in the variable elimination order, and step by step construct a sequence of ATs increasing in size. Initialize $\mathcal{C}^{|V|} = \{\{j_{|V|}\}\}$, $\mathcal{A}^{|V|} = \emptyset$, $L^{|V|} = \emptyset$. Iteratively for $k = |V| - 1, \cdots, 1$ construct $\mathcal{T}^k = (\mathcal{C}^k \cup \mathcal{A}^k, L^k)$ from $\mathcal{T}^{k+1} = (\mathcal{C}^{k+1} \cup \mathcal{A}^{k+1}, L^{k+1})$:

  - Case 1: If $D_k \in \mathcal{A}^{k+1}$: Attach the new clique $D_k \cup \{j_k\}$ to the existing almond D_k, i.e.
    - $\mathcal{A}^k = \mathcal{A}^{k+1}$
    - $\mathcal{C}^k = \mathcal{C}^{k+1} \cup \{D_k \cup \{j_k\}\}$
    - $L^k = L^{k+1} \cup \{\{D_k, D_k \cup \{j_k\}\}\}$
  - Case 2: Else-If $D_k \subset A \in \mathcal{A}^{k+1}$ and A is the smallest such almond: Attach the new clique $D_k \cup \{j_k\}$ to the existing almond A via a new smaller almond D_k, i.e.
    - $\mathcal{A}^k = \mathcal{A}^{k+1} \cup \{D_k\}$
    - $\mathcal{C}^k = \mathcal{C}^{k+1} \cup \{D_k \cup \{j_k\}\}$
    - $L^k = L^{k+1} \cup \{\{D_k, D_k \cup \{j_k\}\}, \{D_k, A\}\}$
  - Case 3: Else-If $D_k \in \mathcal{C}^{k+1}$: Augment the existing clique $\mathcal{C}^{k+1}$ to also include j_k, i.e.
    - $\mathcal{A}^k = \mathcal{A}^{k+1}$
    - $\mathcal{C}^k = \left(\mathcal{C}^{k+1} \cup \{D_k \cup \{j_k\}\}\right) \setminus \{D_k\}$
    - $L^k = \{e \in L^{k+1} \ : \ D_k \notin e\} \cup \left\{\{A, D_k \cup \{j_k\}\} \ : \ \{A, D_k\} \in L^{k+1} \right\}$
  - Case 4: Else: Find the clique $C \in \mathcal{C}^{k+1}$ with $D_k \subset C$: Attach the new clique $D_k \cup \{j_k\}$ to the existing node C via a new almond D_k:
    - $\mathcal{A}^k = \mathcal{A}^{k+1} \cup \{D_k\}$
    - $\mathcal{C}^k = \mathcal{C}^{k+1} \cup \{D_k \cup \{j_k\}\}$
    - $L^k = L^{k+1} \cup \{\{D_k, D_k \cup \{j_k\}\}, \{D_k, C\}\}$

- Present $\mathcal{T} = \mathcal{T}^1 = (\mathcal{C}^1 \cup \mathcal{A}^1, L^1)$ as the constructed AT.

The different options for add-ons are visualized in Figure 3.6.

Note that these ATs have edges between clique nodes and almond nodes as well as between pair of almond nodes in general, i.e. $L \subset \{\ \{N, A\}\ :\ N \in \mathcal{A} \cup \mathcal{C}, A \in \mathcal{A}\}$. An example BN together with its corresponding AT is visualized in Figure 3.7. Also note, as for standard JTs, that there is a clique node in the AT for each LPD table from the BN such that each variable in the LPD domain is represented in the clique.

The final part of the initialization of the AT is to associate each node $N \in \mathcal{C} \cup \mathcal{A}$ with a table D_N. The assignment of values to the tables follows the same procedure as for standard JTs, where the tables for the almond nodes in the AT are treated as the tables for the separators (edges) in the standard JT, and the tables for the clique nodes in the AT are treated as the tables for the nodes in the standard JT. Each almond node $A \in \mathcal{A}$ is given a table D_A of 1s, i.e. the D_A-entry reads 1 for all assignments to the variables X_A. As for standard JTs, the product over the tables of the cliques is required to equal the product over all LPDs. This can be ensured by initializing all clique tables D_C as a table of 1s (as for the almonds), and subsequent for each LPD updating the table of a clique containing the domain variables to be the product of itself and the LPD.

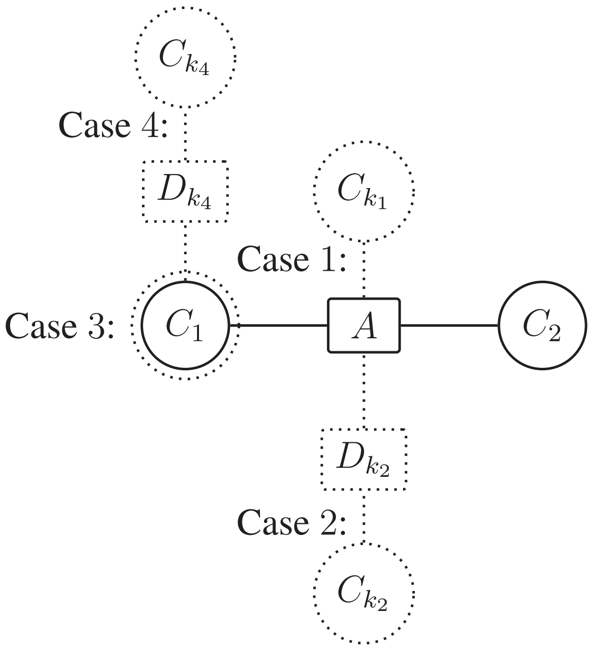

Figure 3.6: Options 1-4 in the AT construction algorithm adds on an existing smaller AT. Here the existing AT is represented with clique nodes C_1 and C_2 and almond node A.

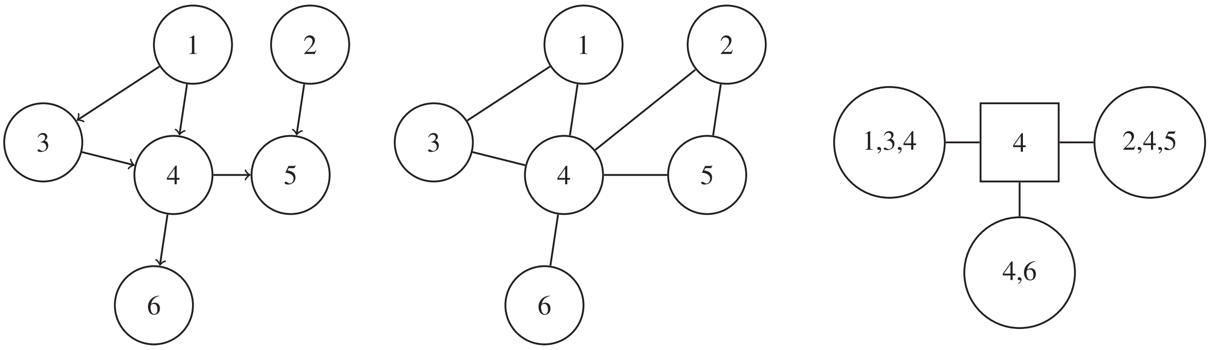

Figure 3.7: A BN on the left, with its moral graph in the middle and corresponding AT on the right. Note that the moral graph is triangulated, as both cycles are of length three.

The next section describes how the AT structure is used to evaluate a query $\mathbb{P}\left(X_R|X_B\right)$ for a BN.

3.6 Message passing in ATs

After the construction and initialization of the AT above, we have $\mathbb{P}\left(X_V\right) = \prod_{C\in\mathcal{C}} D_C$. In fact,

$$\mathbb{P}\left(X_V\right) = Z \cdot \frac{\prod_{C\in\mathcal{C}} D_C/Z_C}{\prod_{A\in\mathcal{A}} {D_A}^{m(A)}}, \qquad Z = \prod_{C\in\mathcal{C}} Z_C \tag{3.1}$$

with unknown normalization constant Z_C for each clique table D_C, is the working assumption of the message passing in JTA. The power $m(A)$ equals the multiplicity of the almond as defined in Chapter 3.5. Before the first message passing, the working assumption (3.1) is fulfilled since

- For each $i \in V$: $\exists! C \in \mathcal{C}$ where D_C is a product of $P(X_i|X_{\mathrm{Pa}(i)})$ and potentially other

LPDs

- $Z = 1$ (since each $Z_C = 1$, as known from (2.2))
- For each $C \in \mathcal{C}$: D_C is a product $\prod_{i \in C'} P(X_i | X_{\text{Pa}(i)})$ for a set $C' \subset C$ where also $\text{Pa}(C') \subset C$
- For each $A \in \mathcal{A}$: D_A evaluates to 1 for each assignment to the variables X_A.

In fact, each sweep of the message passing routine assumes the starting point is of the form in (3.1), and at the end point the form (3.1) is kept with:

- $Z_C = 1 \, \forall C \in \mathcal{C}$, hence $Z = 1$ (the distribution is normalized) and the previous normalizing constant is reported.
- Given $N \in \mathcal{C}$ or $N \in \mathcal{A}$, D_N is the marginal distribution of X_N according to the joint distribution (3.1) before message passing.

We run the message passing immediately after the initialization without any evidence instantiated to obtain permanent values for the representation

$$\mathbb{P}\left(X_V\right) = \frac{\prod_{C \in \mathcal{C}} D_C^0}{\prod_{A \in \mathcal{A}} {D_A^0}^{m(A)}} = \frac{\prod_{C \in \mathcal{C}} \mathbb{P}\left(X_C\right)}{\prod_{A \in \mathcal{A}} \mathbb{P}\left(X_A\right)^{m(A)}}. \tag{3.2}$$

The values of $\{D_C^0\}_{C \in \mathcal{C}}$ and $\{D_A^0\}_{A \in \mathcal{A}}$ are used as the initial starting point for each query. That is, for each query $\mathbb{P}\left(X_{R_k} | X_{B_k}\right)$, the message passing is run from this state (3.2) with an intermediate step of instantiating the evidence B_k.

As for standard JTs, the actual message passing relies on a choice of root in the AT, as the message passing first goes sequentially from each AT node towards the root and then in the opposite order back again. Note that for ATs, both clique nodes and almond nodes are represented in the ordering. A message should leave each AT node after receiving messages from all its neighbors which are further away from the root. Assume an ordering of the AT nodes $N_1, \cdots N_m$ such that for each node's assigned index the further-out-from-root neighbors each have a larger index. That is, we a let the chosen root be N_1, and index the other AT nodes such that $\forall j \in \{1, \cdots, m\}$ the (unique) path $(N_{k_1}, N_{k_2}, \cdots, N_{k_s})$ from N_j to N_1 has $j = k_1 > k_2 > \cdots > k_s = 1$. A possible choice which fulfils the message order requirement is to number the AT nodes in the order they were constructed in the AT construction algorithm above. Note that an augmentation of a clique (Case 3) does not count as a new clique construction for the numbering. This numbering and choice of root ensures that for a pair of almond nodes which are neighbors, the larger one is closest to the root. We will assume this property in the following, as it simplifies the first sweep in the message passing algorithm. For a given ordering, define a function $r : \mathcal{C} \cup \mathcal{A} \to \mathcal{C} \cup \mathcal{A}$ such that $r(N)$ is the (unique) neighbor of N closest to the root, similarly as for standard JTs.

Assume a query of the form $\mathbb{P}\left(X_R | X_B\right)$. As for the VE-algorithm as well as message passing in standard JTs, we have to instantiate the evidence X_B. As for standard JT message passing, it is sufficient to insert the evidence to one clique for each variable: For each variable X_b, $b \in B$, find a clique C containing b and set to 0 all entries in D_C violating the evidence assignment to X_b. After instantiating the evidence, the unknown Z according to (3.1) is exactly $\mathbb{P}\left(X_B\right)$. For simplicity, update each almond table D_A entry to $1/(D_A)^{m(A)}$, such that $\mathbb{P}(X_V) \propto \prod_{N \in \mathcal{C} \cup \mathcal{A}} D_N$ with unknown normalization constant $\mathbb{P}(X_B)$.

Recall that every AT node is either a subset or a superset of its neighbor, for any of its neighbors. There could be both almond nodes and clique nodes among the neighbors of an almond node. The clique nodes only have almond neighbors, which are subsets of the clique. Message passing in the AT then simplifies to:

- **Message passing towards root from node N in order $N_m, \cdots N_2$:** Construct a table D from the table D_N which has the domain of $D_{r(N)}$, i.e.
  - If $r(N) \supset N$, we pass a *message from almond N to larger almond or clique $r(N)$:* Let D have an entry for each possible assignment $X_{r(N)} = x_{r(N)}$ to $X_{r(N)}$, and let its value $D(x_{r(N)})$ be the value of D_N in the entry $D_N(x_N)$ corresponding to the assignment $X_N = x_N$ restricted to the variables in N. Now, D has several identically valued entries and carries all of the information in D_N. Update the values of D_N to be 1 for all entries.
  - Else, we pass a *message from clique N to almond $r(N)$:* Let D have an entry for each assignment to the vector $X_{r(N)}$, such that D is a smaller table than D_N. For each assignment $X_{r(N)} = x_{r(N)}$, let its value be the sum of the values of D_N for assignments $X_N = x_N$ not violating $X_{r(N)} = x_{r(N)}$, i.e. $D(x_{r(N)}) = \sum_{X_{N \setminus r(N)} = x_{N \setminus r(N)}} D_N(x_N)$. Further, divide each entry of D_N used to calculate this sum by the value of the sum.
  - Update $D_{r(N)}$ to be the entry-wise product of itself and the newly constructed D entry-wise: $D_{r(N)}[x_{r(N)}] \leftarrow D_{r(N)}[x_{r(N)}] \cdot D[x_{r(N)}]$.
- **Intermediate processing :**
  - Calculate $Z = \sum_{X_{N_1}} D_{N_1}(X_{N_1})$
  - Report $\mathbb{P}(X_B) = Z$
  - Update $D_{N_1}(x_{N_1}) \leftarrow D_{N_1}(x_{N_1})/Z$, such that the full distribution $\prod_{N \in \mathcal{C} \cup \mathcal{A}} D_N$ is normalized
- **Message passing outwards from root from node N in order $N_1, \cdots N_{m-1}$:** Node N replies to all incoming messages:
  - For each $N_j \in \text{Ne}(N) \setminus \{r(N)\}$ (the nodes from which there was a towards-root-message to N), multiply an appropriate version of the table D_N to the table D_{N_j} by creating a larger or smaller version as in the message passing towards root. Note that now there is no updating of D_N.
- **Result:** $\mathbb{P}(X_N | X_B) = D_N$ for any node $N \in \mathcal{C} \cup A$, almond or clique node. Also, the right side of (3.2) is now a representation of the distribution of X_V conditional on the evidence X_B with all normalization constants $Z_C = 1$ (and consequently $Z = 1$). As for message passing in standard JTs, this formula also holds for all sub-ATs for which all leaves are clique nodes.

To compare with the VE-algorithm and standard JTA for the out-star with n binary variables (left side Figure 3.1), we let the single almond $\{1\}$ be the root. Each of the $n-1$ messages towards root is calculated with time-complexity 2^2, and the root collects with a total time complexity $2n$, and calculates the normalizing constant and normalizes in total time-complexity $2+2$. Each of the messages from root are exactly the normalized potential in the root, so no calculations for the outgoing message are needed. Finally, each of the $n-1$ clique nodes collects with time complexity $2 \cdot 2^2$, and we end up with a total time-complexity of $14n-8$, again linear in n and presenting (conditional) marginals for all cliques and separators. We even gained extra efficiency compared to the JTA with a standard JT in this very special case due to a single AT almond node compared to $n-1$ standard JT separators and the clever choice of root in the AT.

3.7 Time complexity

As the VE-algorithm does not guide the elimination sequence, one cannot expect an efficient sequence. However, no time is spent evaluating different elimination sequences. In general, one must be prepared for the worst case time complexity of $\mathcal{O}(2^n)$ for a joint distribution with n binary variables. According to Lauritzen & Spiegelhalter (1988), the time-complexity of message passing for the JTA with standard JTs is $\mathcal{O}(2^\gamma \cdot K + g \cdot \Theta)$, where

- $K = \sum_{N \in \mathcal{N}} |\Omega_N|$, referred to as the total state space (Ω_N is the state space of the variables in N, such that $|\Omega_N| = 2^{|N|}$ when N has only binary BN-variables),
- $g = |\mathcal{N}|$, the number of nodes in the JT,
- $\gamma = \sup_{N \in \mathcal{N}} |N|$, the maximal number of BN-nodes in a JT-node, and
- $\Theta = \sup_{N \in \mathcal{N}} |\Omega_N|$, the largest state space of a clique.

Obviously, $K \leq g \cdot \Theta$. Also for binary BN-variables, $\Theta = 2^\gamma$. We see that it is preferable to have few BN-nodes in each JT-node, as the expression is linear in g and exponential in γ. Recall that finding the optimal JT is $\mathcal{NP}$-complete due to the triangulation step, so in applications a heuristic is used to try to find a sufficiently good JT.

The time-complexity of JTA is the same based on ATs as it is based for standard JTs, as the almond nodes do not have a significant effect in general. However, for some BNs, we can utilize the AT-structure. We summarize the time-complexity for the out-star with n binary variables in Table 3.1.

Table 3.1: Time complexity for the three variations of algorithms for marginalizing a joint distribution for an out-star with n BN-nodes. For the two versions of the JTA, the construction of the JT/AT is not included.

VE	JTA w JT	JTA w AT
$(n+2) \cdot 2^n - 2$	$26n-2$	$14n-8$
Exponential	Linear	Linear

4 Information Criteria

In several decision problems, it is useful to collect additional information. Then, a set of new questions emerges. What information is worth collecting? Which information is more informative? Are the sources of information correlated? Which combination of tests are the best? Or, in which sequence should we perform different tests?

Figure 4.1 visualizes an area in the North Sea where it could be interesting to search for hydrocarbons. Due to planning and seasonal constraints, a set of m drilling sites must be selected for the initial exploration phase. How should we compare different sets of drilling sites? How do we evaluate the amount of information in an observation? The BN is originally from Martinelli et al. (2011), and will be discussed further in Lilleborge et al. (2016a) and Lilleborge & Eidsvik (2015).

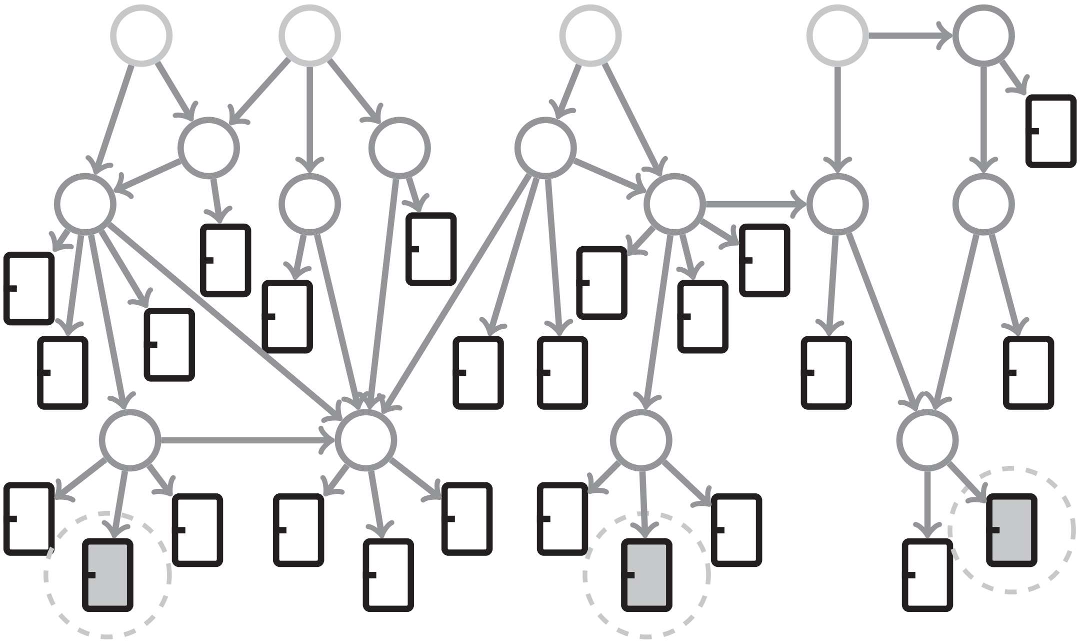

Figure 4.1: A BN representing an oil exploration area in the North Sea. Each root node (light gray) represents a smaller area where hydrocarbons might have been created, and the leaf nodes (black doors) represent potential drilling sites. The rest of the network contains a representation of the potential migration paths (dark gray), which together with the root nodes build the correlation structure between the variables of interest, the potential drilling sites, visualized as doors one could choose to open. A possible choice of 3 potential drilling sites for initial exploration is marked with light gray filling and a dashed circle. The information value of this set is evaluated by an information measure, and compared with the information measure of other candidate observation sets before any observations are made.

Design of experiment is often associated with research about the procedure for assigning treatment to subjects, most notably Fisher (1935). According to Box et al. (2005), one should "block what you can and randomize what you cannot" when dealing with unavoidable sources of variability, while “hard thinking” is required otherwise. While randomization is a general procedure for eliminating systematic differences between treatment and control groups, see Gerber & Green (2012), few would recommend collecting random information. When selecting between sources of information, thoughtful evaluation is necessary. In the next section, a tool for analysing information gathering is presented.

4.1 Value of Information

Value of Information (VoI) is a way of evaluating the value of additional information for a given decision problem. This decision theory concept allows for comparison between different types of future data gatherings, by evaluating their impact on the result of the final decision through probabilistic inference. That is, we are in a setting where we are to make a decision, like a medical doctor evaluating whether a patient should undergo cancer treatment or not. Say, we are to decide on an action a from the set of possible actions $\mathcal{A}$. The outcome of action a depends on the outcome of a random variable X, and has value $u(a, x)$ in the case of $X = x$. The function u is referred to as the utility function, and its value $u(a, x)$ for a given action a and outcome x is referred to as the utility of a and $X = x$. That is, the utility $u(a, x)$ represents an evaluation of the usefulness or how valuable the outcome x is for the decision maker after taking action a, see Hamburg (1970) for a further discussion.

It is optimal to choose the action a that maximizes the expected value $\mathbb{E}_{[X]}u(a, X)$, and the prior value of the decision problem is defined as

$$\text{PV} = \max_{a \in \mathcal{A}} \left\{ \mathbb{E}_{[X]} u(X, a) \right\},$$

Before the final decision is to be made, we are given the option to do one or several tests $t \in \mathcal{T}$, but of course, the different tests all have associated costs. For decision-making processes in for instance medicine, it is very important to do the right tests and the right amount of tests before the decision is made. In finance, the choice of additional information can influence the expected profit. In applications, information is not perfect as the data includes noise and potentially the data could also be incomplete. Because smoking is a risk factor for cancer, information about a patient’s smoking habits can help his doctor estimate the patients risk for lung cancer, but it is not sufficient information to ensure a correct diagnosis. The radiologist might be uncertain about how to interpret the findings on a mammogram also after additional imaging (UL, MR) is taken into account. That is, we need to model the uncertainty or variability of the test results. We can model the test as another variable Y which is correlated with X. After the value of Y is known, we expect a posterior value

$$\text{PoV} = \mathbb{E}_{[Y]} \max_{a \in \mathcal{A}} \left\{ \mathbb{E}_{[X|Y]} u(X, a) \right\}$$

The total VoI of observing Y is therefore given by the expected increase in value, i.e. the difference between posterior value and prior value

$$\text{VoI}(Y) = \mathbb{E}_{[Y]} \max_{a \in \mathcal{A}} \left\{ \mathbb{E}_{[X|Y]} u(X, a) \right\} - \max_{a \in \mathcal{A}} \left\{ \mathbb{E}_{[X]} u(X, a) \right\}$$

We say it is rational to pay up to the amount of $\text{VoI}(Y)$ to observe the realization of Y.

VoI quantifies the value of different data sources, as it takes into account the different outcomes of the data gathering in a probabilistic framework. The VoI framework allows for questions like how much information one should gather, and provides a formula for what a given set of data is worth for the given application. Tests could also be taken simultaneously or sequentially. A medical doctor might order a biopsy if the ultrasound is positive. Obviously, we expect new information to reduce the uncertainty in a problem, and VoI analysis further ensures that it is suboptimal to pay anything for information that cannot change the final decision. Eidsvik et al. (2015) provide an introduction to VoI for applications like energy, geophysics, geology, mining, and environmental science.

4.2 Static or sequential selection

After performing a given test, it might still be optimal to perform another. The medical doctor might order MR, ultrasound, blood work and a variety of other tests before he has enough confidence to set a diagnosis. What would be the next test could depend on the result of the current test. The problem of finding the optimal test sequence taking the results into account is commonly solved by a technique called dynamic programming. Dynamic programming is described in Cormen et al. (2009), and dynamic programming for the petroleum exploration case (e.g. Figure 4.1) is thoroughly explained in Martinelli et al. (2013). In other situations, one has to select a static set of tests before any of them is performed, due to planning or seasonal constraints.

Assume you want to find the subset $B_m \subset K$ of size m which maximizes a set function $v()$. To solve the problem exactly, you would have to compare the $\binom{|K|}{m}$ possible subsets of size m. A Greedy[1] Forward Search only evaluates parts of the subspace by iteratively solving m smaller subproblems

$$\tilde{B}_0 = \emptyset, \tilde{B}_i = B_{i-1} \cup \arg \min_{\{j\} \,:\, j \in K \setminus B_{i-1}} v(\{j\} \cup B_{i-1}).$$

Note that this approach solves a forward sequential selection problem in order to find a reasonable solution to the static problem. Also observe that $|K| + 1 - i$ subsets are compared in the ith level, and $\frac{1}{2}m(2|K| - m + 1)$ subsets are evaluated in total. Thus, the Greedy Search provides a fast approximate solution to the optimization problem. For some other problems, a Greedy Search is proved to provide the optimal solution. An obvious example mentioned in the previous chapter, is Prim's algorithm for a minimum spanning tree. If v is submodular, Krause & Guestrin (2005) prove that the Greedy Forward Search solution for subset selection comes with a constant factor approximation guarantee.

Similarly, one can define a Greedy Backward Search by initializing $\tilde{B}_{|K|} = K$ and removing one element at a time. The Backward Search is not close to being as popular as the Forward Search in practice, as evaluation of $v(B)$ usually has complexity increasing heavily with $|B|$ in many applications.

[1] Also referred to as One Step Look Ahead

4.3 Information and ethics

Not all decision problems have utilities naturally measured in monetary value, like expected profit in dollars. In medicine we can set a price on a test representing the equipment and labor spent on performing the test. In addition, it is often also a price for the patient in the form of discomfort, stress or even pain which can make one type of test preferable to another. However, the hardest quantity to specify is usually a value to human life or survival. One may ask if it is ethical to say that a patient's life is not worth more than $\$X$, but without such a limit it would be rational e.g. for a country to spend the whole national budget on trying to avoid cancer in the population. Further, money saved on more effective testing or treatment can be spent on improving the health care service. In Lilleborge et al. (2016b), we use VoI to calculate the optimal breast cancer screening program based on data from $200,000$ Norwegian women, and propose that "resources saved by cutting suboptimal testing in low risk groups may justify spending more resources on high risk groups".

4.4 Information and measures of information

Parameter estimation problems obtain information about a parameter from data from the probability distribution in question. One can then ask how much information a data sample can provide about the unknown parameter. Quantifying the amount of information allows us to compare different experiments. In decision theory, information is obtained in order to choose a better action and improve profits. The VoI criterion allows us to choose the experiment with highest expected profit. In some situations the profit is not naturally measured in monetary value. There are several ways of measuring information in terms of evaluating the effect of learning in terms of probability updates. The selection of a measure should depend on the statistical model, but most importantly the current application. Information measures are important in design of experiment, as they help evaluate which data are most informative. In this section, we follow the notation of Ginebra (2007).

If the realization of a variable X in our experiment has a large probability according to our prior belief, we can think of it as confirming and requiring only a small update of our belief. The new information introduced by the observation of X is then small. On the other hand, when the realization of X is of low probability and requiring a larger update of our belief, we can assign the realization of X a larger information value. Similarly as in Maximum Likelihood analysis, we consider $\ell(x|\theta) = \log f(x|\theta)$ and associate $\frac{\partial}{\partial\theta}\ell(x|\theta)$ close to 0 as in indication of θ having a value that assigns high probability to the realization of X. Fisher (1922) introduces the Fisher information about the parameter θ in an experiment where the random variable X is realized from the pdf f,

$$I(\theta) = \mathbb{E}_{[X|\theta]}\left[\frac{\partial}{\partial\theta}\ell(X|\theta)\right]^2 = \mathbb{V}\mathrm{ar}_{[X|\theta]}\left[\frac{\partial}{\partial\theta}\ell(X|\theta)\right] = -\mathbb{E}_{[X|\theta]}\left[\frac{\partial^2}{\partial\theta^2}\ell(X|\theta)\right]. \tag{4.1}$$

The equalities are obtained between the tree different expressions of $I(\theta)$ under the assumption that one can interchange derivation and integration for $\mathbb{E}_{[X|\theta]}\frac{\partial}{\partial\theta}f(x|\theta)$ and $\mathbb{E}_{[X|\theta]}\frac{\partial^2}{\partial\theta^2}f(x|\theta)$.

Similarly when θ is a k-dimensional vector, $\left[\frac{\partial}{\partial\theta}\ell(X|\theta)\right]_i = \frac{\partial\ell(X|\theta)}{\partial\theta_i}$ and $\left[\frac{\partial^2\ell(X|\theta)}{\partial\theta^2}\right]_{i,j} = \frac{\partial^2\ell(X|\theta)}{\partial\theta_i\partial\theta_j}$, and we obtain the information matrices

$$I(\theta) = \mathbb{E}_{[X|\theta]}\left[\frac{\partial\ell(X|\theta)}{\partial\theta}\left(\frac{\partial\ell(X|\theta)}{\partial\theta}\right)^T\right] = \mathrm{Cov}\left[\frac{\partial\ell(X|\theta)}{\partial\theta}\right] = -\mathbb{E}_{[X|\theta]}\left[\frac{\partial^2\ell(X|\theta)}{\partial\theta^2}\right]. \tag{4.2}$$

Lindley (1956) discusses Shannon's Information Entropy $-\mathbb{E}_{[\theta]}\log\mathbb{P}\left(\theta\right)$ for a parameter θ, and studies the expected change in value for this quantity for a given experiment. This definition of entropy was first introduced by Shannon (1948) in the context of information theory in communications engineering. Lindley (1956) interprets the information in a prior distribution by looking at how much information must be provided before the value of θ is known. Among other properties, he establishes that one always expects to gain information, $\Delta_X = -\mathbb{E}_{[\theta]}\log\mathbb{P}\left(\theta\right) - \left(-\mathbb{E}_{[X]}\mathbb{E}_{[\theta]}\log\mathbb{P}\left(\theta|X\right)\right) \geq 0$, but the gained information of two experiments are not necessarily additive $\Delta_{Y,X} \leq \Delta_Y + \Delta_X$. Shannon Entropy is probably the most used information criteria, and a few application areas are medicine in Westover et al. (2012), piezometric data in Bueso et al. (1998) and sulfate concentration records in Ko et al. (1995).

Fisher (1922) and Lindley (1956) both argue for their way of measuring information by ensuring that their measure has good properties. Blackwell (1951), and later Le Cam (1964), provides theoretical discussions comparing two experiments without reference to a measure of choice. In this setting, an experiment can only be preferable to another if one of them is "sufficient" for the other. Otherwise, they are not comparable. Kullback & Leibler (1951) introduce the Kullback-Leibler divergence $\int_X \frac{dP}{dQ}dP$ of a distribution Q from a distribution P as a relative entropy inspired by Shannon (1948). Csiszár (1967) uses this to introduce a more general concept of divergence.

Ginebra (2007) studies what in general can qualify as a information measure for an experiment. This places the specific information measures of Fisher (1922) and Lindley (1956) in a generous class adhering to the rigorous general theoretic considerations of Blackwell (1951). We are in the setting of $E = \{(X, S_X); (P_\theta, \Omega)\}$ being a very general statistical experiment. We observe a random variable $X \in S_X$ which has an unknown distribution P_θ among the possible distributions $\{P_\theta\}_{\theta\in\Omega}$ of X. However, as our focus in Lilleborge et al. (2016a), Ginebra (2007) does comparison "made on statistical merit only, irrespective of experimental costs". He argues that an information measure $I(E)$ should assign a real number to each experiment, it should evaluate to 0 if there is no learning (no probability updates), and it should prefer experiment E to experiment F ($I(E) \geq I(F)$) if E is at least as good as F for any terminal decision problem. These three are his minimal set of requirements for a measure of information.

He concludes that finding the most informative experiment can be expressed as a decision problem with the following characteristics:

- The utility function is convex.
- The reward of a given experiment is the likelihood ratio or posterior distribution statistic of the outcome.
- The information of an experiment is its expected utility.

- Choose the experiment that maximizes information.

Assume there are k options for θ, $\Omega = \{\theta_1, \cdots, \theta_k\}$, and choose positive $\{\pi_i\}_{i=1}^k$ which ensures that the convex combination $P_\pi = \sum_{i=1}^k \pi_i P_{\theta_i}$ dominates each P_θ. Let K_π be the convex hull of $\{(1/\pi_1, \cdots, 0), \cdots, (0, \cdots, 1/\pi_k)$. Define as a minimum sufficient characteristic of the statistical properties of E (through its distribution),

$$T : S_X \to K_\pi \qquad \text{s.t.} \qquad T_\pi(X) = \frac{1}{p_\pi(X)} \left(p_{\theta_1}(X), \cdots, p_{\theta_k}(X)\right).$$

Ginebra argues that "the sufficiency principle dictates that the information has to be measured through functions of $T_\pi(x)$ and common wisdom dictates that these functions have to be such that the further $T_\pi(x)$ is away from $(1, \cdots, 1)$ towards an extreme point of K_θ, the larger values they take." He concludes that the generalized ϕ-divergence measure of the information about θ in a realization $X = x$ from an experiment is $\phi\left(T_\pi(x)\right)$ for a non-negative convex $\phi(u)$ with $\phi(1, \cdots, 1) = 0$, and interprets it as a measure of the surprise about θ in X. Further, Ginebra gives interpretations of several well-known and much used information measures, such as the Shannon Entropy, in the light of the theory presented.

The choice of information measure should depend on the application. In the oil exploration case (see Figure 4.1) treated in the Lilleborge et al. (2016a) and Lilleborge & Eidsvik (2015), each potential drilling site will eventually be drilled or not drilled, and this decision is made individually for each potential drilling site X_i based on the probability of finding hydrocarbons $\mathbb{P}\left(X_i = 1\right)$. In the case of no drilling, the probability of success will not be explored. The information measure should strive to minimize the variability of each potential drilling site (minimize or maximize probability of hydrocarbons), in order for the decision maker to be as certain as possible about the decision of drilling or not for each potential drilling site. Thus, it is natural to select an information measure that minimizes a sum of individual expected variability-evaluations for each site. For example, this could be $\mu_1(B) = \mathbb{E}_{[X_B]}\left[\sum_{i=1}^L \text{Var}(X_i|X_B)\right]$ where we sum over the collection of potential drilling sites L and the expectation is taken over an initial exploration observation set $B \subset L$. As the variance of a Boolean variable is largest for $p = \frac{1}{2}$, minimizing this measure means striving to get probability updates $\mathbb{P}\left(X_i = 1|X_B\right)$ away from $\frac{1}{2}$. A simple transformation of the conditional probabilities $\mathbb{P}\left(X_i|X_B\right)$ can be used to manipulate the measure to prefer updates away from e.g. the critical probability p_c which makes the decision maker indecisive. In applications where all variables will be explored and we care about learning the number of successes, a version like $\mu_2(B) = \mathbb{E}_{[X_B]}\left[\text{Var}\left(\sum_{i=1}^L X_i|X_B\right)\right]$ is more appropriate. As $\mu_2(B) = \mu_1(B) + 2\sum_{i<j\in L} \mathbb{E}_{[X_B]}\left[\text{Cov}(X_i, X_j|X_B)\right]$, we see that the aim of getting a stable estimate of the sum explicitly results in a penalty for positive covariances and equally weighted benefit for negative covariances.

The theory of information measures has some links to, but should not be confused with, InfoQ introduced in Kenett & Shmueli (2014). InfoQ is a quite general information quality concept considering a goal g (e.g. causal explanation, prediction, descriptive statistics or tests), some data X, an empirical analysis method f (e.g. statistical parametric/semiparametric/non-parametric models, data mining etcetera) and a utility measure U (e.g. predictive accuracy, goodness-of-fit, statistical power). The InfoQ is defined by InfoQ$(f, X, g) = U[f(X|g)]$ (see

Kenett & Shmueli (2014)) and is constructed to evaluate the potential of "a particular dataset to achieve a particular goal using a given empirical analysis method".

In Lilleborge et al. (2016a), we discuss information measures for applications similar to the oil exploration case in Figure 4.1. Lilleborge & Eidsvik (2015) provide an algorithm constructing converging upper and lower bounds to efficiently select the optimal observation set according to a given information measure. If the BN distribution is MTP_2, Lilleborge (2016) presents a tailored lower bound for a more efficient search. Lilleborge et al. (2016b) use VoI analysis to analyze the optimality of the Norwegian Breast Cancer Screening Programme.

5 Aims of Thesis

This thesis is about BNs, a highly active research area. However, unlike most other recent works on BNs, this thesis is not about building the network but rather on how to utilize an already built model. I assume that the BN is known; a given structure consisting of a graph and corresponding parameters learnt from data, expert knowledge or a combination. The aims of this thesis are built upon exploration of the following question:

> Given a BN defined by expert knowledge and/or data,
> which observations should be made to gain maximum information?

Gabriele Martinelli's thesis provided background knowledge on information gathering to maximize the expected profit of dependent prospects in an oil exploration problem where a collection of prospects is selected for drilling (i.e. Martinelli et al. (2011), Martinelli et al. (2012), Martinelli et al. (2013)). However, maximizing profit in this setting is highly dependent the future oil price and future development costs. Moreover, Martinelli focus on dynamic strategies, while rig constraints and drilling seasons requires the drilling campaign to be planned as a static group.

To maximize information gain, one needs to understand how to measure information. This requires a study of information criteria for BNs. What has made the Shannon Entropy so popular, and how should this quantity be interpreted? Which other criteria are used for various applications? Which properties should one require from information criteria in general, and how should one select an appropriate information criterion? Which criteria are best suited for the oil exploration case?

After selecting an appropriate information criterion, the optimization still remains. How should one ensure maximum gain of information for a given criteria? As for dynamic sequential oil exploration, the statical subset selection problem is expected to have high time-complexity. Can the probabilistic structure of the BN model be utilized for fast structured optimization? Can attributes from JTA be cleverly applied in the optimization? Also, Martinelli & Eidsvik (2014) studied clustering strategies, but the question of how to build an efficient optimization algorithm was left open.

Aiming for a more purely statistical approach to maximal information gain, the resulting theory should be general enough to be applied to different application areas. In addition to the petroleum exploration case which initiated this project, an application within medicine is interesting and appropriate to show applicability of the thesis.

In summary, the three main aims of this thesis are:

1. Explore information criteria for BNs and non-sequential exploration designs for BNs
2. Fast structured optimization of information criteria for subset selection

3. Show applicability of general theory by applying it to two different application areas, namely petroleum exploration and medicine.

6 Summary of papers

6.1 Paper I

LILLEBORGE, M., HAUGE, R. & EIDSVIK, J. (2016a). Information Gathering in Bayesian Networks Applied to Petroleum Prospecting. ***Mathematical Geosciences*** **48, 233–257**

The value of information approach with a monetary utility function is usually the most natural information measure whenever costs and revenues for the underlying decision problem are well known. In many contexts it is not easy to associate appropriate cost and income functions to the decision problem; in other cases one chooses a best estimate. In these cases it can be appropriate to apply purely information based measures. This paper explores different criteria for efficient information gathering and for optimal design of BNs.

Lilleborge et al. (2016a) study criteria which allow for comparison of the information based on probabilistic merits only. This might be necessary whenever the costs and/or possible gains depend on quantities which are highly unknown. In oil exploration, the future price for oil is such an unknown parameter where the estimated value has a large influence on the optimal decision. An alternative approach is to aim for maximal reduction of the total uncertainty. The information criteria in this paper are calculated as a function of the probability distribution alone. Each criterion looks for observations or tests that give information about more than the few variables we are observing in each such test. The information measures we studied assign values based on correlations and conditional dependence structures in the BN. Each information criteria discussed is related to Ginebra (2007), which provides general theory for properties of information measures.

It is important to understand what each information measure is expressing and why the data collection is carried out. The paper is discussing differences and similarities of the different measures. Different properties means the measures are tailored for different approaches or applications. This again means the choice of information measure should be highly dependent on the application. In this paper we consider a set L of observable variables from which a subset of variables should be chosen for observation. The different measures discussed evaluates the total remaining uncertainty in all variables of L.

The Shannon Entropy-measure is well known and successfully applied in several applications. In our setting, the Shannon entropy criteria chooses the observation set associated with the highest uncertainty in itself, without consideration to probability updates in the unobserved variables. This property is clearly undesirable in a setting where information criteria are used to guide learning about several correlated variables (also the unobserved ones), like in the petroleum exploration case. Lilleborge et al. (2016a) further provide guidelines for choosing an information measure in applications similar to the petroleum exploration case, where one cares about each of the observable variables after the selected observations are made.

6.2 Paper II

LILLEBORGE, M. & EIDSVIK, J. (2015). **Efficient designs for Bayesian networks with sub-tree bounds.** ***Statistics and Computing*** **, To appear** The information measures discussed in Lilleborge et al. (2016a) all have time-complexity exponential in the size of the observation set B. Further, the search for the optimal observation set $B^\star \subset K$ of size m has $\binom{|K|}{m}$ possible candidates. (The number of candidates is of order $|K|^m$ hence also exponential in the size of B for $m << |K|$.) Solving the optimization problem by comparing the values of each candidate, lead us to focus on small observation set sizes in Lilleborge et al. (2016a). In this paper we look for fast structured optimization of information gain.

This paper aims to tackle the high time-complexity by use of upper and lower bounds. The paper describes the construction of upper and lower bounds such that they can be iteratively improved, and the resulting sequence of bounds is converging to the true information values. The converging bounds are applied in a search strategy where the candidate set is reduced iteratively as the bounds ensure that candidates are suboptimal. This way, we ensure that the algorithm returns the true optimal candidate. This algorithm can also be stopped after a given amount of time or after reaching a given threshold for a guarantee, and the current best candidate is presented together with a percentage guarantee of its value compared to the (unknown) true optimum. We also describe how the converging bounds can be applied in established fast approximation schemes like a greedy search or an exchange algorithm to further accelerate these algorithms.

Similarly as in Martinelli & Eidsvik (2014), we use clairvoyant information and clustering strategies to construct the bounds. In Martinelli & Eidsvik (2014), the network is divided into disjoint clusters. For each cluster, the Markov blanket is analyzed manually to find appropriate variables for clairvoyant information. For information measures, the clairvoyant information results in a lower bound, while probability updates restricted to be from variables within each local cluster results in an upper bound. In this paper, we utilize the JT constructed for JTA to automatically find appropriate clairvoyant variables. It turns out the separators in the JT are efficient choices of clairvoyant variables. By removing the restriction of the clusters to be disjoint, and replacing it with a unique local cluster for each variable, we end up with a construction that intuitively allows for iterative improvements of the bounds: As the clairvoyant separators are further out in the graph, and we include all variables in a sub-graph within a boundary of clairvoyant separators for exact probability updates, the bounds are approaching the true measure values.

In the paper, we compare the results for the true optimum and the approximations for the North Sea network as well as for some simulated examples. The run-times of the different strategies are also compared. For small m (i.e. where it is available), we also present run-time of naive optimization by comparison of measure values. By the tables of run-times, we see that the converging bounds search for the true optimal candidate has clear reductions in run time compared with the naive exact calculations, and this allows us to tackle larger problem sizes. Obviously, the established approximation schemes result in much better time-complexity, with a trade-off of no guarantee of optimality.

6.3 Paper III

LILLEBORGE, M. (2016). Efficient optimization with Junction Tree bounds in discrete MTP2 distributions. Tech. rep., Norwegian Computing Center

This technical report presents methodology which was developed but never finished for publication during my time as a PhD student. The work was initiated as we realized that the time complexity encountered in the calculations for Lilleborge et al. (2016a) was limiting our scope for observation set sizes m. I was familiar with optimizing a set function by upper and lower bounds through my master thesis, and started studying the messages in the JTA to look for patterns or ways of approximating these in a controlled way. This lead to the idea of a strongest possible message from different directions in a JT, and the MTP$_2$ concept allowed for combination of strongest possible messages from different directions. For simplicity, we focus on binary random variables.

If a discrete random variable has an MTP$_2$ distribution, we say that it is positively associated and we have Cov$(f(X), g(X)) \geq 0$ for any functions g and f. Let $\vee$ and $\wedge$ denote the operator on two vectors which returns a vector of the entry-wise maximum and minimum, respectively. A distribution is MTP$_2$ if $\mathbb{P}(X) \cdot \mathbb{P}(Y) \leq \mathbb{P}(X \vee Y)\mathbb{P}(X \wedge Y)$ for all X, Y in its support. Assume the random vector has binary variables as entries. The MTP$_2$ assumption introduces a rule for which assignments of some variables which would maximize the conditional success probability of another, since a success always increase the success probability of all other variables.

The JT groups the variables according to probabilistic dependencies such that updates from BN nodes in one JT node propagates to a non-neighboring JT node through the probabilistic updates for the intermediate JT-nodes on the unique path between them. An altering of a distribution of a JT node introduces an altering of the distribution of the neighbor, and the altering of the distribution of the neighbor introduces an altering of the distribution of a further out neighbor in the JT, and so on.

Combining the streamlined updating pattern in the JTA with the uniform covariance pattern of MTP$_2$ distributions, we construct converging upper and lower bounds for information measures using local calculations in the JT similar to the bounds in Lilleborge & Eidsvik (2015). For the special case of MTP$_2$ distributions, these bounds will be faster to calculate, and in some cases they will be tighter. However, they will need some extra pre-processing together with the initialization step of the JTA.

As I have not encountered a large enough BN with the MTP$_2$ property to motivate optimization through bounds nor data to construct such a network, the theory has not been published. The more general clustering and clairvoyance bounds were applied on the North Sea case and several other simulated BNs in Lilleborge & Eidsvik (2015).

6.4 Paper IV

Lilleborge, M., Hofvind, S., Sebuødegård, S. & Hauge, R. (2016b). Using Bayesian Networks to optimize performance of the Norwegian Breast Cancer Screening Program - a modelling study. ***Submitted for publication in Statistics in Medicine***

In this paper, we apply knowledge from the earlier works of this thesis to breast cancer screening. First, a graphical model is used to estimate cancer risk based on results of the previous screening test and self-reported information about risk factors such as lifestyle and family history of breast cancer. Secondly, we implement this cancer risk in an estimated BN where the true cancer status is represented together with current screening test results. Finally, we provide a value of information analysis to optimize for the best test regime.

This paper provides a theoretical mathematical evaluation of the optimal performance of a breast cancer screening program, and aims to contribute towards the possibility of improving the efficiency of the Norwegian Breast Cancer Screening-Program. The work tries to answer a highly relevant question of today, according to the following recent encouragements from breast cancer research:

> Recommendations about the frequency of mammography should be personalized on the basis of a woman's age, breast density, history of breast biopsy, and family history of breast cancer, as well as the effect of mammography on her quality of life (conclusion of paper; Schousboe et al. (2011))
>
> The time has come for individualized screening (quote from review paper; Desreux et al. (2012)).

Today, the mammograms of all participants of the Norwegian Breast Cancer Screening Program are double-read and all women are screened every 2 years. The two independent radiologists each give a score $1-5$ on the images. If a woman gets at least one score at level 2 or higher, the two radiologists meet for a consensus where they decide if the woman gets a recall letter for additional imaging and possibly a biopsy. There is no stratification based on breast cancer risk in the current program.

Our model defines four breast cancer risk groups (low risk(17.6%), middle$\div$ risk(69.2%), middle$+$ risk(12.4%) and high risk(0.76%)) based on age and results of the previous screening mammogram. For the low risk group it is sufficient to do screening every 4 years. For the other risk groups, the screening mammograms should first be single-read. A second independent interpretation should be done if the highest score from the first radiologist is on level $2-3$ (middle$\div$ risk), $1-3$ (middle$+$ risk) or $1-2$ (high risk), respectively. For higher scores the woman should be referred to additional imaging, and for a score at level 1 a woman in the middle$\div$ risk group should be evaluated as cancer free by the single-read mammogram. The paper further discusses bounds for cancer risk levels for when the two radiologists should have a consensus meeting, as well as lower risk bounds for additional imaging and biopsy after previous tests.

7 Discussion

As discussed in Chapter 1, BNs are commonly used in a wide range of applications. However, there have been limited contributions from statisticians in the design of experiments for these models and for decision-making. It is certainly important to study these higher-level tasks, to bring statistics closer to policy-making. Most existing research on BNs consider the problem of how to build the network; from observational data, from expert knowledge or a combination of both. The focus in the four papers of this thesis is rather on how to apply the information in the network.

The question of maximum information gain is one of decision analysis, and theoretical works like Ginebra (2007) provide general theory about information measures. Besides discussions of properties for a selection of information measures, the first paper Lilleborge et al. (2016a) provides guidelines for selection of information criteria. Different information criteria can recommend dissimilar strategies, so the final decision might be determined by the selection of information criteria. Also, a criteria which has been successfully applied numerous times, might give undesirable results in a different setting. The take home message is that one needs to consider the application to know why the information criterion is used, and from that evaluate which criteria are applicable.

The wide flexibility of the BN models leaves a wealth of opportunities, but it also leaves us to deal with a model where learning and information evaluation might seem less intuitive than for a less complex model. It turns out it is important to be aware of this flexibility when an information criteria is selected. One of the contributions of Lilleborge et al. (2016a) is that a less desirable property of the Shannon Entropy for oil exploration and similar applications can be much more dominating for BNs than for more uniform models like e.g. spatial statistics models or Gaussian random field models.

For large graphs and large observation sets, the time-complexity of the naive optimization of information criteria introduces a need for more efficient algorithms to find the optimal observation set. The second paper Lilleborge & Eidsvik (2015) and the third paper Lilleborge (2016) contribute on optimization schemes for doing well in a large graph where the exact solution is not tractable due to exponential growth of the solution space and enormous storage problems on the computer. By thorough understanding of message passing and the general structure of the JTA, properties of the joint distribution can be utilized by simple computations rather than running JTA as a black box repeatedly until all possibilities are evaluated. For complex calculations, it is important to comprehend when enough is understood as well as avoid re-computing the same quantities over and over.

The approaches of Lilleborge & Eidsvik (2015) and Lilleborge (2016) have similarities, but are built on two different ways of studying the JTA. The bounds of Lilleborge & Eidsvik (2015) are constructed based on how the JT orders the BN variables according to dependence structure.

Further, the MTP$_2$-tailored lower bound of Lilleborge (2016) is constructed on information propagation in the message passing. In addition to presentation of its bounds and the different algorithms, Lilleborge & Eidsvik (2015) provide background information about the JTA with illustrations in the appendix.

By exploring attributes from the JTA, the algorithm in Lilleborge & Eidsvik (2015) tailors a clustering approximation to the computational structure of the JTA. The algorithm does not make any assumptions about the covariance structure of the variables and provides a run-time reduction in sparse BN models. Some level of sparsity is a common assumption for large BN models; a growing model incorporates more variables and hence more edges, but the density of edges is often assumed to be bounded. However, with BN models both the sparsity pattern induced by the edges and the covariance pattern induced by the parameters can vary throughout the network according to data.

The main contribution to the run-time reduction in Lilleborge & Eidsvik (2015) is due to computations on local subparts of the JT. The upper and lower bounds do not consider all possible assignments of the observation set B at all evaluations, but it considers additional variables as well. The relatively small problem sizes ($33, 42$ and 117 nodes, respectively) of the networks presented in Lilleborge & Eidsvik (2015) do not bring out the full potential of the algorithm. However, we did not have the computer power nor data to analyse much larger networks.

The MTP$_2$-tailored bound in Lilleborge (2016) requires the distribution of the observable variables to have the MTP$_2$-property. This is obviously a special case, but it allowed for applying intuition about message passing in JTA to construct another type of bound.

The JTA is today established as the standard inference engine for BNs. Implementations of the JTA are easily accessible, and this allows for repeatedly calling this routine blindly without reference to the structure behind the calculations. This has allowed for easy calculations in complex distributions, and has obviously been an important resource in many applications. However, Lilleborge & Eidsvik (2015) and Lilleborge (2016) have illustrated how insights into JTA can increase efficiency of how the JTA is used. This suggests that several of the many diverse applications where the JTA is used today might benefit from specializing their algorithm to their use.

To bring the deep level understanding of probability propagation in BNs to a practical application, real world data from the Cancer Registry was analyzed in the fourth paper Lilleborge et al. (2016b) together with domain experts. The analysis proposes a more efficient breast cancer screening program stratified by an estimated breast cancer risk model.

7.1 Future work

As mentioned in Lilleborge et al. (2016a), applications where the observable nodes and the scoring nodes are disjoint or partially overlapping are left as future work. Further, towards the end of Chapter 4, we mentioned the difference between the measures in Lilleborge et al. (2016a) of the form $\sum_i \mathbb{E}_{[X_B]} f(\mathbb{P}(X_i|X_B))$ and more portfolio-based versions of the form $\mathbb{E}_{[X_B]} f\left(\sum_i \mathbb{P}(X_i|X_B)\right)$. A future study of the latter types of measures could be interesting.

Applications of the bounds of Lilleborge & Eidsvik (2015) to larger networks is something I would appreciate. The general idea of the construction of the bounds should leave a wealth of opportunities for applications for where the theory can be applied. The technical report Lilleborge (2016) is still premature, but I certainly hope it gets an opportunity to evolve towards a more attractive state.

The time-complexity of the naive optimization of information gain is a product of two exponential factors exponential in the size m of the observation set, the first representing the calculation of a measure value and the second representing the number of candidates. The computation of the bounds aims to tackle the first factor (faster calculation of measure value). I discussed branch and bound and other optimization algorithms with prof. Geir Dahl at UiO, who has expert knowledge about optimization, but we ended up concluding that to tackle the second factor, we had turn to approximation schemes. Note that if the information measure is submodular, the optimization problem can be efficiently (meaning with polynomial time-complexity) solved, see Schrijver (2000). To do exact and efficient optimization for information gain is still an open question as long as the information measure is not submodular.

Lilleborge et al. (2016b) provide a mathematical analysis of breast cancer screening, and evaluate a possibility of improving the efficiency of the Norwegian Breast Cancer screening-program. The analysis provides an important evaluation of the program, as well as instructing a different way of analyzing the program for the Cancer Registry of Norway and other organizers of similar screening programs. However, the breast cancer risk model built for this study does not include all well-known risk factors for breast cancer, and is built with the R-package “gRim” of Højsgaard (2012). Obviously, there are many breast cancer risk models in the literature, usually based on a Cox-model. These models tend to focus on how the risk develops over longer risk horizons, like five years, ten years and life-time risks. For our study, we learned a graphical model from anonymized data to predict the risk at a given screening round conditional on risk factors as well as the results from the previous screening round two years earlier. A more carefully evaluated discriminative breast cancer risk model with more risk factors could utilize the value of information analysis results further, and possibly result in a better stratified screening recommendation.

References

ALMOND, R. & KONG, A. (1991). Optimality issues in constructing a markov tree from graphical models. Tech. rep., Department of Statistics, Harvard University.

BLACKWELL, D. (1951). Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, Calif.: University of California Press.

BONDY, J.-A. & MURTY, U. S. R. (2008). *Graph theory*. Graduate texts in mathematics. New York, London: Springer.

BOX, G. E. P., HUNTER, J. S. & HUNTER, W. G. (2005). *Statistics for experimenters : design, innovation, and discovery*. Wiley series in probability and statistics. Hoboken (N.J.): Wiley-Interscience.

BUESO, M., ANGULO, J. & ALONSO, F. (1998). A State-Space Model approach to Optimum Spatial Sampling Design based on Entropy. *Environmental and Ecological Statistics* **5**, 29–44.

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

CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L. & STEIN, C. (2009). *Introduction to Algorithms*. The MIT Press, 3rd ed.

COWELL, R., DAWID, P., LAURITZEN, S. & SPIEGELHALTER, D. (2007). *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Statistics for Engineering and Information Science Series. Springer.

CSISZÁR, I. (1967). Information-type measures of difference of probability distributions, and indirect observations. *Studia Scientiarum Mathematicarum Hungarica* **2**, 229âĂŞ318.

DESREUX, J., BLERET, V. & LIFRANGE, E. (2012). Should we individualize breast cancer screening? *Maturitas* **73**, 202 – 205.

DURRETT, R. (2007). *Random Graph Dynamics*. Cambridge: Cambridge University Press.

EIDSVIK, J., MUKERJI, T. & BHATTACHARJYA, D. (2015). *Value of Information in the Earth Sciences: Integrating Spatial Modeling and Decision Analysis*. Cambridge University Press.

FISHER, R. (1935). *The design of experiments. 1935*. Edinburgh: Oliver and Boyd.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **222**, 309–368.

GERBER, A. & GREEN, D. (2012). *Field Experiments: Design, Analysis, and Interpretation*. W. W. Norton.

GINEBRA, J. (2007). On the Measure of the Information in a Statistical Experiment. *Bayesian Analysis* **2**, 167–212.

HAMBURG, M. (1970). *Statistical analysis for decision making*. Harbrace series in business and economics. Harcourt, Brace & World.

HØJSGAARD, S., EDWARDS, D. & LAURITZEN, S. (2012). *Graphical Models with R*. Use R! Boston: Springer.

HØJSGAARD, S. (2012). Graphical independence networks with the gRain package for R. *Journal of Statistical Software* **46**, 1–26.

JENSEN, F. V. & NIELSEN, T. D. (2007). *Bayesian Networks and Decision Graphs*. Springer Publishing Company, Incorporated, 2nd ed.

JORDAN, M. I. (t.a.). Conditional independence and factorization. In *An Introduction to Probabilistic Graphical Models*. To appear.

KENETT, R. S. & SHMUELI, G. (2014). On information quality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **177**, 3–38.

KO, C. W., LEE, J. & QUEYRANNE, M. (1995). An Exact Algorithm for Maximum Entropy Sampling. *Operations Research* **43**, 684–691.

KOLLER, D. & FRIEDMAN, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

KRAUSE, A. & GUESTRIN, C. (2005). Near-optimal value of information in graphical models. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

KULLBACK, S. & LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

LAURITZEN, S. L. & SPIEGELHALTER, D. J. (1988). Local Computation with Probabilities on Graphical Structures and their Application to Expert Systems (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **50**, 157–224.

LE CAM, L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Statist.* **35**, 1419–1455.

LILLEBORGE, M. (2016). Efficient optimization with Junction Tree bounds in discrete MTP2 distributions. Tech. rep., Norwegian Computing Center.

LILLEBORGE, M. & EIDSVIK, J. (2015). Efficient designs for Bayesian networks with sub-tree bounds. *Statistics and Computing* , To appear.

LILLEBORGE, M., HAUGE, R. & EIDSVIK, J. (2016a). Information Gathering in Bayesian Networks Applied to Petroleum Prospecting. *Mathematical Geosciences* **48**, 233–257.

LILLEBORGE, M., HOFVIND, S., SEBUØDEGÅRD, S. & HAUGE, R. (2016b). Using Bayesian Networks to optimize performance of the Norwegian Breast Cancer Screening Program - a modelling study. *Submitted for publication in Statistics in Medicine* .

LINDLEY, D. V. (1956). On a Measure of the Information provided by an Experiment. *Annals of Mathematical Statistics* **27**, 986–1005.

MARTINELLI, G. & EIDSVIK, J. (2014). Dynamic Exploration Designs for Graphical Models using Clustering with Applications to Petroleum Exploration . *Knowledge-Based Systems* **58**, 113–126.

MARTINELLI, G., EIDSVIK, J. & HAUGE, R. (2013). Dynamic Decision Making for Graphical Models applied to Oil Exploration. *European Journal of Operational Research* **230**, 688–702.

MARTINELLI, G., EIDSVIK, J., HAUGE, R. & FØRLAND, M. D. (2011). Bayesian Networks for Prospect Analysis in the North Sea. *AAPG Bulletin* **95**, 1423–1442.

MARTINELLI, G., EIDSVIK, J., HAUGE, R. & HOKSTAD, K. (2012). Strategies for petroleum exploration based on bayesian networks: a case study. In *SPE Annual Technical Conference and Exhibition, SPE 159722*.

POURRET, O., NAÏM, P. & MARCOT, B. (2008). *Bayesian Networks: A Practical Guide to Applications*. Statistics in Practice. Wiley.

RUSSELL, S. & NORVIG, P. (2003). *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2nd ed.

SCHOUSBOE, J. T., KERLIKOWSKE, K., LOH, A. & CUMMINGS, S. R. (2011). Personalizing mammography by breast density and other risk factors for breast cancer: Analysis of health benefits and cost-effectiveness. *Annals of Internal Medicine* **155**, 10–20.

SCHRIJVER, A. (2000). A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *J. Comb. Theory Ser. B* **80**, 346–355.

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423.

WESTOVER, M. B., EISEMAN, N. A., CASH, S. S. & BIANCHI, M. T. (2012). Information theoretic quantification of diagnostic uncertainty. *The Open Medical Informatics Journal* .

YANNAKAKIS, M. (1981). Computing the minimum fill-in is np-complete. *SIAM Journal on Algebraic Discrete Methods* **2**, 77–79.

II

Title	**Efficient optimization with Junction Tree bounds in discrete MTP$_2$ distributions**
Author	**Marie Lilleborge** `<Lilleborge@nr.no>`
Quality assurance	Ragnar Hauge, Assistant Research Director SAND
Date	May 25, 2016
Publication number	SAND/07/2016

Abstract

This paper construct upper and lower bounds for expected values of convex functions for Multivariate Totally Positive distributions of order 2. The bounds can be iteratively improved, and can be used to optimize information measures or value of information utilities for Bayesian Networks with the property of being Multivariate Totally Positive of order 2. The bounds are applied on a simple illustrating synthetic example with binary variables for simplicity.

The construction of the bounds is inspired by the local updates made by the message passing in the Junction Tree Algorithm, and utilizes the coherent covariance structure of Multivariate Totally Positive distributions of order 2. The resulting formula allows for preprocessing in the full joint to construct pairs of smaller local distributions covering different parts of the network, from which the upper and lower bounds are efficiently constructed.

Keywords	MTP2, Bayesian Network, Upper and Lower Bounds, Subset selection, Design of experiment, Value Of Information, Information measures, Almond Tree, Junction Tree Algorithm
Target group	Anyone
Availability	Open
Project	
Project number	
Research field	Statistics
Number of pages	26
© Copyright	Norwegian Computing Center

1 Introduction

Upper and lower bounds can be used to find the exact optimal solution of problems which are infeasible through comparison of exact values. Bounds are also a way to approximately solve complex problems with a guarantee, as bounds for the difference between the value of an approximate best candidate and the value of the true optimum follows. In this paper, we focus on subset selection for experimental design. The general design problem has been extensively studied in the statistics literature, but here we focus on design of experiments for graphical models like Bayesian networks (BNs).

The Junction Tree Algorithm (JTA) for BN probability assessments initially developed by Lauritzen and Spiegelhalter (1988) is commonly considered the most efficient algorithm for calculating probability statements for BNs. The JTA has an initial step constructing a computational object called a Junction Tree, and then for each probability assessment this object is looped over twice in a message passing routine. This paper is the result of research on the JTA which ended up being divided into two branches, one published in Lilleborge and Eidsvik (2015) and the other resulting in this technical report. The similar assumptions as in Lilleborge and Eidsvik (2015) are as follows: We assume a BN with node set V and random variables in a vector $X_V = [X_v]_{v \in V}$. A subset $K \subseteq V$ of the nodes are associated with the observable variables X_K, while $X_{V \setminus K}$ are latent variables which help specify the full probability model. The BN examples considered in this paper have binary random variables taking values in $\{0, 1\}$. We focus on the observation set selection problem, i.e. finding the observation set $B \subset K, |B| = m$ of size m which optimizes the information measure selected for the application. The JT of the BN is central, and the reader is referred to Lilleborge and Eidsvik (2015) for background on BNs and the JTA, on information measures and on optimization with upper and lower bounds.

As in Lilleborge and Eidsvik (2015), the lower bound constructed in this paper will allow for the message passings to happen on a smaller subset of the JT. As the bounds are converging to the true value, this subset is increased to the full JT. When optimizing over many candidates, this allows for removing elements from the candidate set iteratively as the bounds are improved. Hopefully, the size of the candidate set is increasing more rapidly than the computational complexity of the bounds. This strategy was successful for exact solution as well as for faster algorithms in Lilleborge and Eidsvik (2015).

Our focus in this paper is on static designs for BNs, just as in Lilleborge and Eidsvik (2015), trying to tackle the time-complexity issues of optimizing subset selection for information measures, see e.g. Lilleborge et al. (2015). The static design problem consider a set of variables of which an optimal subset should be selected, i.e. selecting the most informative sample according to an information measure. As is common in experimental design, the goal is to select a subset of nodes for experimentation, with no opportunities for adaptive selection. The reader is referred to Peyrard et al. (2013) and Bonneau et al. (2014) for adaptive (sequential) sampling designs for graphs or Markov random fields. Closer to the approach in this paper, Brown and

Smith (2013) and Martinelli and Eidsvik (2014) evaluates adaptive designs for BNs by use of bounds for the sequential selection of sites. We suggest that the lower bound constructed in this paper also could be applied to the adaptive sequential sampling problem if the distribution of the variables is MTP$_2$.

The difference between the theory presented here and the one discussed in Lilleborge and Eidsvik (2015), is the MTP$_2$ assumption. Where the theory in Lilleborge and Eidsvik (2015) is based on how the variables are arranged according to their covariance pattern in a JT, the theory of this paper is based on the message passing. The message passing is the system of probability updates in the JTA, directing how information is distributed by messages or signals between the JT nodes. The JT groups the variables according to probabilistic dependencies, and places variables that are more correlated closer to each other. In fact, it constructs chains of groups of variables such that the outermost groups are dependent due to a mutual dependence to the intermediate groups. These chains are all appearing in a tree-structure, the JT. Further, updates from BN nodes in one JT node propagates to a non-neighboring JT node through the probabilistic updates for the intermediate JT-nodes on the unique path between them. This can be described sequentially for an observation of variables which appear in the same JT node C: First, observing variables in JT node C, gives local updates for the marginal distribution for X_C. As the JT node knows its marginal distribution, this can be done locally. A neighboring JT node N^1 contains some of the same variables as C, and it is obvious that the marginal of X_{N^1} needs to be updated so that the marginal of $X_{C\cap N^1}$ is the same in both JT nodes. In fact, this exactly the update that is needed, and the distribution of X_{N^1} can be updated to $\mathbb{P}\left(X_{N^1}\right) = \mathbb{P}\left(X_{C\cap N^1}\right) \cdot \mathbb{P}\left(X_{C\setminus N^1} | X_{C\cap N^1}\right)$ where the first factor is calculated from the updated marginal in the JT C and the second factor is calculated from the un-touched marginal in the JT N^1. A further neighbor $N^2 \neq C$ of N^1 will get updates through the distribution of N^1 through the same procedure, and so on. When the distribution of a JT node is updated, it introduces updates for the distribution of the neighbor, and the updates for the distribution of the neighbor introduces an update of the distribution of a further out neighbor in the JT, and so on.

In this paper, the MTP$_2$ assumptions allows for a tailored lower bound, as the assumption of a MTP$_2$ distribution defines a unified rule for the assignment of each observable node which correspond to the strongest signals to the other observable nodes. This will be discussed in Section 4. Success propagating tree-networks and Naive Bayes models (the latter possibly requiring a re-labelling of states) are examples of distributions with the MTP$_2$ property. The North Sea network studied in Lilleborge et al. (2015) and Lilleborge and Eidsvik (2015), however, possess so-called explaining-away effects via intermediate nodes with multiple parents. Assume X_1, X_2, X_3 are binary variables, X_1 and X_2 are independent and $\mathbb{P}\left(X_3 = 0 | X_1, X_2\right) = (1-p)^{X_1+X_2}$ for some $p > 0$. Note that an observation of a success in X_3 increases the success probability of X_1, but a subsequent observation of a success in X_2 would again decrease the success probability of X_1. The first increase is due to $X_1 = 1$ being a good explanation for $X_3 = 1$, however as $X_2 = 1$ is an equally good explanation, this later evidence is used to "explain away" the first. Explaining-away effects are incompatible with MTP$_2$. As explaining-away effects are present in the North Sea Network of Lilleborge and Eidsvik (2015) and Lilleborge et al. (2015), the theory presented in this technical report cannot be applied to provide a lower bound for information value for this North Sea network. However, informal test runs verify that the bound actually serves as a good approximation strategy in this case.

In the following, the probability of the event $\{A = a\}$, that the outcome of a Random Variable A has value a, is denoted by $\mathbb{P}(A)$. That is, we let the assignment be implicit. This simplified notation makes intuitively sense in this paper since we are not concerned about the actual outcome a but the expected value of a function f of the distribution of A, as in $\mathbb{E}_{[A]}f(A)$. We do not include the assignment of the random variables because it will be integrated out, as the upper and lower bounds are constructed as expected values. Thus, the function evaluation $f(X_R = x_R)$ is referred to as $f(X_R)$ also for the random vector $X_R = [X_i]_{i\in R}$. The expected value of the function $f()$ applied on a vector X_R of binary variables $X_i,\ i \in R$ is defined as

$$\mathbb{E}_{[X_R]}f(X_R) = \sum_{X_R=x_R\in\{0,1\}^{|R|}} f(X_R)\mathbb{P}(X_R)\,.$$

Similarly, the conditional expectation of one variable X_i (given an assignment to X_R) is

$$\mathbb{E}_{[X_i|X_R]}f(X_{R\cup\{i\}}) = \sum_{X_i=x\in\{0,1\}} f(X_{R\cup\{i\}})\mathbb{P}(X_i|X_R)\,.$$

As the evaluation of each design must be done before the variables are actually observed, an information measure consists of an inner function expectation with respect to the distribution conditional on the assignment of an observation set and an outer expectation where the conditional assignment is finally integrated out.

The paper is structured as follows. In Section 2, the MTP$_2$ assumption is defined and discussed. In Section 3 the upper and lower bound of Lilleborge and Eidsvik (2015) are defined. The MTP$_2$ lower bound is defined in Section 4. The upper bound of Lilleborge and Eidsvik (2015) can be applied together with the MTP$_2$ lower bound constructed in this paper, and I will compare the lower bound of Lilleborge and Eidsvik (2015) to the MTP$_2$ lower bound. In Section 5, all three bounds are applied to a synthetic BN example. Finally, Section 6 provides closing remarks.

2 Total positivity

In this work, we assume a type of positive dependence between the variables of interest X_K. From Fallat et al. (2016) we find that the random vector X_K is said to be positively associated if $\text{Cov}(f(X_K), g(X_K)) \geq 0$ for any non-decreasing functions f and g. They also add that all known definitions of positive dependence are implied by something called the MTP$_2$ constraints, as follows:

Definition 1. *A random vector $X \in \chi$ is Multivariate Totally Positive of order 2 (MTP$_2$) if its density function* p *fulfils*

$$p(x)p(y) \leq p(x \wedge y)p(x \vee y) \qquad \forall x, y \in \chi. \tag{1}$$

The main purpose of Fallat et al. (2016) is to prove that if a probability distribution is MTP$_2$ and has coodinate-wise connected support, then it is faithful to its concentration graph. In the following, however, the MTP$_2$ property will be used to construct upper and lower bounds for a set function. Some other useful statements from Fallat et al. (2016) about a MTP$_2$ random variable X_K are

1. The MTP$_2$ property is closed under conditioning and marginalization; i.e. for $B \subset K$ both $X_B|X_{K\setminus B} = x_{K\setminus B}$ (for a.e. $x_{K\setminus B}$) and X_B are MTP$_2$.

2. For any subset $B \subset K$ and non-decreasing function ϕ, the conditional expectation $\mathbb{E}_{[X_B]}\phi(X_B|X_{K\setminus B} = x_{K\setminus B})$ is non-decreasing in $x_{K\setminus B}$.

3. For a decomposable graph $\mathcal{G}$ such that the intersection of any two cliques are either empty or a singleton, a distribution $\mathbb{P}(\cdot)$ which is Markov with respect to $\mathcal{G}$ is MTP$_2$ if and only if the marginal distribution of each clique is MTP$_2$.

The first property ensures that MTP$_2$ for all variables implies MTP$_2$ for a smaller collection, and follows easily from the definition. The second property is an important part of the proof for the bounds we will later define, and the third property allows us check if a larger distribution is MTP$_2$ part by part. (See Fallat et al. (2016) for proofs.)

2.1 A single parent network

As a simple example of a MTP$_2$ distribution, we will study a Bayesian Network with a single parent with N children. For $N = 3$ we are in the situation of Figure 1. From the previous section, we know that it is sufficient to check each parent P - child C pair to ensure the MTP$_2$

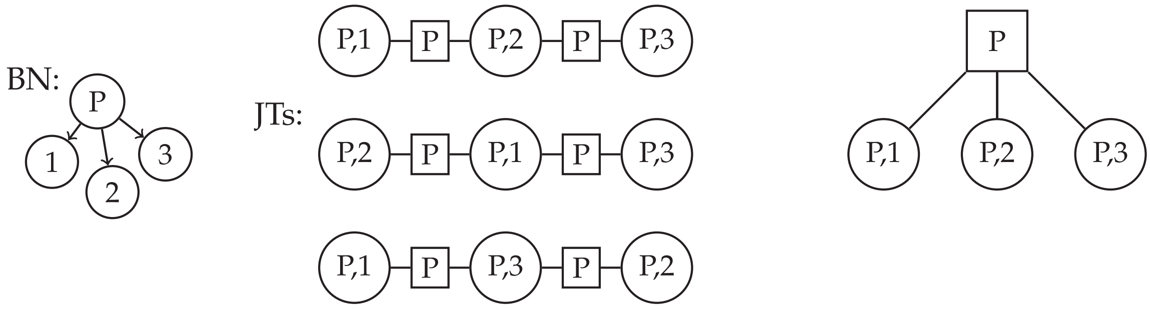

Figure 1. A BN with one parent and three children (left picture) has four possible Junction Trees (JT) (right pictures). However, all three middle configurations have two equal separators, and by merging equal separators as in Almond and Kong (1991) all three JTs result in the simpler JT on the right.

property. Obviously from (1), it is sufficient to check

$$\mathbb{P}\left(P=0,C=1\right)\cdot\mathbb{P}\left(P=1,C=0\right)\leq\mathbb{P}\left(P=1,C=1\right)\cdot\mathbb{P}\left(P=0,C=0\right),$$

which reduces to $\mathbb{P}\left(C=1|P=0\right)\leq\mathbb{P}\left(C=1|P=1\right)$ if P is not deterministic. Note that if P is in fact deterministic, the above equation is always fulfilled as the children are independent (and we will ignore this case). That is, a single parent binary network is MTP$_2$ if and only if a success for the parent increases the success probability for each child.

3 Upper and lower bounds from by clustering and clairvoyance

As in Lilleborge and Eidsvik (2015), assume our set of observable nodes K is a disjoint union of clusters C from a collection $\mathcal{C}$,

$$K=\dot{\cup}_{C\in\mathcal{C}}C.$$

That is, each $C\in\mathcal{C}$ holds a set of observable nodes. The upper and lower bounds will be based on calculations within each cluster; i.e. takes into account the correlations within each cluster, and ignores the correlation between clusters. Since calculations for BNs are performed in a corresponding JT, the partitioning of K into disjoint clusters should be guided by the JT. For example, BN-variables appearing in the same JT-nodes (or some neighboring JT-nodes) can be chosen to be in the same cluster.

By comparing the upper and lower bounds, we are able to give some evaluation to the bounds as approximative values, since each of these errors will be less than the difference between the bounds. Their average will again have half the error bound. We know from Lilleborge and Eidsvik (2015) that for an information measure $\mu_f(B)=\mathbb{E}_{[X_B]}f_T(\mathbb{P}\left(\cdot|X_B\right)$ (f concave), upper and lower bounds can be constructed as

$$\mu_f^U(B)=\mu_f(B^-)\quad\text{and}\quad\mu_f^L(B)=\mu_f(B^+),\qquad B^-\subseteq B\subseteq B^+.$$

Whenever the information measure is of the form

$$\mu_T(B)\equiv\sum_{i\in K}\mathbb{E}_{[X_B]}f_T(\mathbb{P}\left(X_i|X_B\right)),$$

we can regroup the terms to

$$\mu_T(B)=\sum_{C\in\mathcal{C}}\sum_{i\in C}\mu_T^i(B).$$

This is the case for e.g. μ_{Var},μ_{PrE} and μ_{NwE} discussed in Lilleborge et al. (2015), namely the sum of conditional variances

$$\mu_{Var}(B) = \sum_{i \in L} \mathbb{E}_{[X_B]} \left[\mathbb{V}\mathrm{ar}_{[X_i|X_B]} \left[X_i \right] \right],$$

the expected number of prediction errors

$$\mu_{PrE}(B) = \sum_{i \in L} \mathbb{E}_{[X_B]} \left[1 - \max_{x \in \{0,1\}} \left\{ \mathbb{P}\left(X_i = x | X_B \right) \right\} \right],$$

as well as the node-wise sum of entropies

$$\mu_{NwE}(B) = -\sum_{i \in L} \mathbb{E}_{[X_B]} \left[\mathbb{E}_{[X_i|X_B]} \left[\log \mathbb{P}\left(X_i \mid X_B \right) \right] \right].$$

For a given choice of information measure, define the upper bound

$$\hat{\mu}_T(B) \equiv \sum_{C \in \mathcal{C}} \sum_{i \in C} \mu_T^i(B \cap C) = \sum_{C \in \mathcal{C}} \sum_{i \in C} \mathbb{E}_{[X_B]} f_T(\mathbb{P}\left(X_i | X_{B \cap C} \right)),$$

as a version where each node only see probability updates resulting from observations within its own cluster. This is the situation described for the Variance measure in Lilleborge and Eidsvik (2015). From Lilleborge et al. (2015) we know that $\mu_T^i(B \cap C) \geq \mu_T^i(B)$, since less probability updates means less learning. This again ensures that $\hat{\mu}_T(B) \geq \mu_T(B)$. The optimal observation set within a collection $\mathcal{B}$ according to the true ($B^\star$) and upper bound ($\hat{B}$) measure, respectively, are

$$B^\star \equiv \arg\min_{B \in \mathcal{B}} \mu_T(B) \quad \text{and} \quad \hat{B} \equiv \arg\min_{B \in \mathcal{B}} \hat{\mu}_T(B).$$

Note that through the easier-to-calculate $\hat{\mu}_T$ and corresponding minimum $\hat{B}$, we also have an upper bound for the optimum of the true measure $\mu_T(B^\star)$, since

$$\hat{\mu}_T(\hat{B}) \geq \mu_T(\hat{B}) \geq \mu_T(B^\star).$$

To construct a lower bound, we could introduce appropriate clairvoyant information $R = B^+ \setminus B$ for each cluster, namely

$$\tilde{\mu}_T(B) \equiv \sum_{C \in \mathcal{C}} \sum_{i \in C} \mu_T^i(B \cup R(C)).$$

$R(C)$ is some set disjoint from C, for example $R(C) = L \setminus C$. A good choice of $R(C)$ should follow two intuitive requirements. The first requirement (R1), is to select $R(C)$ so that $X_i \perp B \setminus C || R(C)$ for each observable node $i \in C$ in the cluster. This (R1) allows for local sub-JT calculations for each cluster $C \in \mathcal{C}$, as $\mu_T^i(B \cup R(C)) = \mu_T^i((B \cap C) \cup R(C))$ and the latter can be calculated on a sub-JT containing the variables in C and $R(C)$. The second requirement

(R2), aiming for an efficient choice of sub-JT, is to select $R(C)$ so that these variables appear close to C in the full JT. That is, computations for the cluster C needs to happen on a sub-JT containing all variables in both $R(C)$ and C. $R(C)$ "close" to C in the full JT is an intuitive indication that a sub-JT containing all variables in both sets is "small". Also note that smaller argument set $(B \cap C) \cup R(C)$ means less variables to integrate out when calculating μ_T^i, and hence time efficiency. Choosing $R(C) = V \setminus C$, is clearly fulfilling the first point (R1). When the observable nodes are all leaf nodes, this choice is effectively the same as $R(C) = \text{Pa}(C)$ for the collection of BN-parents of each observable node $i \in C$, since $\mu_T^i((B \cap C) \cup (V \setminus C)) = \mu_T^i((B \cap C) \cup \text{Pa}(C))$. The choice ($R(C) = \text{Pa}(C)$) is also adhering to the second point (R2), following from the running intersection property of the JT combined with the fact that the BN-parents of cluster node $i \in C$ must appear together with i in at least one JT-node. For the more general case, the corresponding choice would be to take the union of the markov blankets of each node in C and remove from this set the nodes appearing in C. Actually, a corresponding analysis appears more straightforward in the JT: Select a subtree of the JT in which all nodes in the cluster is represented, and choose as $R(C)$ the separators separating the subtree from the rest of the full JT. This is illustrated in Figure 2. The converging bounds of Lilleborge and Eidsvik (2015) makes a similar initial choice, and iteratively increases the size of the sub-JT for local computations by choices of $R(C)$ further out in the JT.

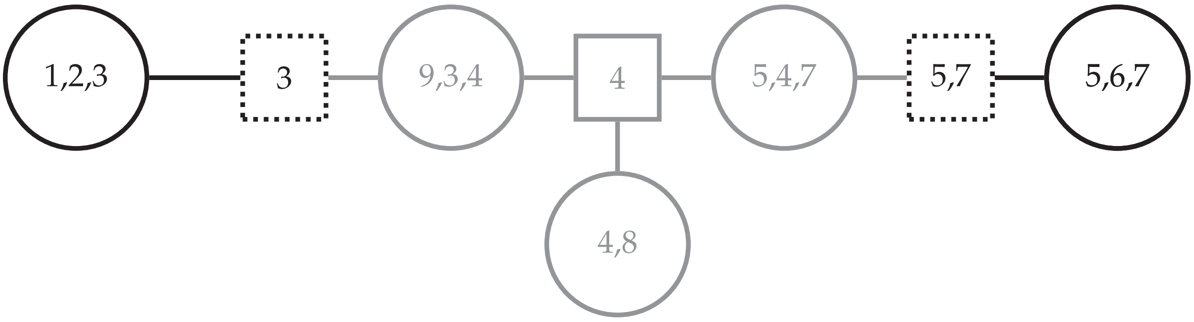

Figure 2. A small JT including 9 BN nodes labelled $1-9$. An example sub-JT for the cluster $\{4,8,9\}$ is marked in grey. From this choice of sub-tree, the R-set is automatically set to $\{3,5,7\}$, namely the variables appearing in the separators marked as dotted square boxes.

4 MTP$_2$ alternative to clairvoyant

Clairvoyant information (e.g. $R(C) = \text{Pa}(C)$) will often correspond to receiving perfect information on nodes $j \notin K$ we only get indications on in practice. Recall that K is the set of observable nodes. Message passing-wise, this corresponds to sending stronger messages (clearer information) in the JTA than for the true measure value of any subset of observable nodes. Instead of creating a lower bound based on stronger messages, we aim to create a lower bound based purely information from observable nodes. That means the strongest indications from observable variables outside the cluster that the cluster variable is 0 or 1, respectively. This could lead to a tighter lower bound than a standard clairvoyant, since there is no information on unobservable nodes more correlated with the cluster. That is, the bound would be tighter if the standard clairvoyant implies perfect information on nodes much more correlated with the variable of interest compared to the observable nodes in the application. It can also result in faster calculations, as preprocessing can give two conditionals to consider compared to the exponentially many assignments of $X_{R(C)}$. What we here refer to as preprocessing is simply storing information otherwise calculated repeatedly by the JTA (i.e. clever implementation of JTA for this case), as only the information within the cluster varies with the observation set B considered. We aim for something comparable to the algorithm in Lilleborge and Eidsvik (2015), where local computations are done on a sub-JT. That algorithm instructs integration over additional information from JT-nodes separating the local cluster from the rest, as variables outside the cluster is conditionally independent of variables within the cluster given the additional information. For the bound constructed in this section, preprocessing or simply clever JTA-runs will lay the groundwork for a lower bound where calculations again are done in a local cluster or a sub-JT. Similarly to Lilleborge and Eidsvik (2015), the sub-JT can be iteratively increased to get a lower bound converging to the true value.

Let $\ddot{i} \in K$ be an observable node in a given cluster $\ddot{C}$. We will focus on the corresponding term $\mu_T^{\ddot{i}}$ to the node $\ddot{i}$ in the full measure $\mu_T(\cdot) = \sum_{i\in K} \mu_T^i(\cdot)$, and assume that all observable nodes are leaf nodes as in Lilleborge and Eidsvik (2015). Using the clairvoyant idea, we assume a thought experiment, where we would observe all observable nodes $K \setminus \ddot{C}$ outside the cluster. However, for computational reasons we only consider the cases $X_{K\setminus\ddot{C}} = x_-,\ X_{K\setminus\ddot{C}} = x_+$ that yields smallest and largest conditional success probability of $X_{\ddot{i}}$, respectively. We will use the reasoning from the thought experiment to consider some observation set $B \subset K$. Similarly to the clairvoyant lower bound, we will add additional information to B from outside of the cluster. That is, in the thought experiment, we consider observations on the corresponding $B \cup \left(K \setminus \ddot{C}\right)$, where $X_{K\setminus\ddot{C}}$ is restricted to $x_\pm$. This will result in a lower bound with terms of the form

$$\begin{aligned}\breve{\mu}_T^{\ddot{i}}(B) \equiv\ & \mathbb{E}_{[X_{B\cap\ddot{C}}]}\left[w^{\ddot{i}}(X_{B\cap\ddot{C}}) \cdot f_T(\mathbb{P}\left(X_{\ddot{i}} = 1|x_-, X_{B\cap\ddot{C}}\right))\right] \\ & + \mathbb{E}_{[X_{B\cap\ddot{C}}]}\left[\left(1 - w^{\ddot{i}}(X_{B\cap\ddot{C}})\right) \cdot f_T(\mathbb{P}\left(X_{\ddot{i}} = 1|x_+, X_{B\cap\ddot{C}}\right))\right], \qquad (2)\end{aligned}$$

where $w^{\ddot{i}}(x_{B\cap\ddot{C}})$ are weights. The measure term $\breve{\mu}_T^{\ddot{i}}$ corresponds to the term of node $\ddot{i}$ in the full measure $\breve{\mu}_T = \sum_{i\in K} \breve{\mu}_T^i$. Note the MTP$_2$ assumption introduces homogeneity which ensures $x_- = \vec{0}$ and $x_+ = \vec{1}$.

4.1 Calculations/Derivation

Let $P \in \mathrm{Pa}(\ddot{C})$ be the parent of the observable node $\ddot{i}$ in cluster $\ddot{C}$, both P and $\ddot{i}$ are binary. Recall that $x_-,\ x_+$ are also the assignments to $X_{K\setminus\ddot{C}}$ that yields smallest and largest conditional success probability of P, respectively. By the MTP$_2$-property,

$$\mathbb{P}\left(P=1|x_-,x_{B\cap\ddot{C}}\right) < \mathbb{P}\left(P=1|x_B\right) < \mathbb{P}\left(P=1|x_+,x_{B\cap\ddot{C}}\right)$$

unless $K\setminus\ddot{C} \perp P \mid B\cap\ddot{C}$. To each $X_B = x_B,\ x_B \in \chi_B$ the restricted assignment $X_{B\cap\ddot{C}} = x_{B\cap\ddot{C}}$ lets us write

$$\exists! t_{x_B} \in [0,1]\ :\quad \mathbb{P}\left(P=1|x_B\right) = t_{x_B}\cdot\mathbb{P}\left(P=1|x_-,x_{B\cap\ddot{C}}\right) + (1-t_{x_B})\cdot\mathbb{P}\left(P=1|x_+,x_{B\cap\ddot{C}}\right).$$

The above equation is equivalent to

$$\mathbb{P}\left(P=0|x_B\right) = t_{x_B}\cdot(1-\mathbb{P}\left(P=1|x_-,x_{B\cap\ddot{C}}\right)) + (1-t_{x_B})\cdot(1-\mathbb{P}\left(P=1|x_+,x_{B\cap\ddot{C}}\right)),$$

which again let us combine and collect terms to see that also

$$\begin{aligned}\mathbb{P}\left(X_{\ddot{i}}=1|x_B\right) &= \sum_P \mathbb{P}\left(X_{\ddot{i}}=1|P\right)\mathbb{P}\left(P|x_B\right) = t_{x_B}[\cdots] + (1-t_{x_B})[\cdots]\\ &= t_{x_B}\cdot\mathbb{P}\left(X_{\ddot{i}}=1|x_-,x_{B\cap\ddot{C}}\right) + (1-t_{x_B})\cdot\mathbb{P}\left(X_{\ddot{i}}=1|x_+,x_{B\cap\ddot{C}}\right),\end{aligned}$$

since $B \perp \ddot{i} \mid P$. For a concave function $f_T(\cdot)$,

$$f_T(\mathbb{P}\left(X_{\ddot{i}}=1|x_B\right)) \geq t_{x_B}\cdot f_T(\mathbb{P}\left(X_{\ddot{i}}=1|x_-,x_{B\cap\ddot{C}}\right)) + (1-t_{x_B})\cdot f_T(\mathbb{P}\left(X_{\ddot{i}}=1|x_+,x_{B\cap\ddot{C}}\right)),$$

and thus

$$\begin{aligned}\mu_T^{\ddot{i}}(B) &\equiv \mathbb{E}_{[X_B]} f_T(\mathbb{P}\left(X_{\ddot{i}}=1|X_B\right))\\ &\geq \mathbb{E}_{[X_B]}\left[t_{X_B}\cdot f_T(\mathbb{P}\left(X_{\ddot{i}}=1|x_-,X_{B\cap\ddot{C}}\right)) + (1-t_{X_B})\cdot f_T(\mathbb{P}\left(X_{\ddot{i}}=1|x_+,X_{B\cap\ddot{C}}\right))\right].\end{aligned}$$

Defining the weights $w^{\ddot{i}}(X_{B\cap\ddot{C}}) = \mathbb{E}_{[X_{B\setminus\ddot{C}}|X_{B\cap\ddot{C}}]} t_{X_B}$, the above calculations prove the lower bound $\breve{\mu}_T^{\ddot{i}}(B) \leq \mu_T^{\ddot{i}}(B)$ for $\breve{\mu}_T^{\ddot{i}}(B)$ defined in (2).

Interpreting $0/0$ as 0, we see that $w^{\ddot{i}}(X_{B\cap\ddot{C}})$ is easily calculated by re-use of quantities we use for other computations:

$$\begin{aligned}w^{\ddot{i}}(X_{B\cap\ddot{C}}) = \mathbb{E}_{[X_{B\setminus\ddot{C}}|X_{B\cap\ddot{C}}]} t_{X_B} &= \mathbb{E}_{[X_{B\setminus\ddot{C}}|X_{B\cap\ddot{C}}]} \frac{\mathbb{P}\left(P=1|x_{B\cap\ddot{C}},x_+\right) - \mathbb{P}\left(P=1|x_B\right)}{\mathbb{P}\left(P=1|x_{B\cap\ddot{C}},x_+\right) - \mathbb{P}\left(P=1|x_{B\cap\ddot{C}},x_-\right)}\\ &= \frac{\mathbb{P}\left(P=1|x_{B\cap\ddot{C}},x_+\right) - \mathbb{P}\left(P=1|x_{B\cap\ddot{C}}\right)}{\mathbb{P}\left(P=1|x_{B\cap\ddot{C}},x_+\right) - \mathbb{P}\left(P=1|x_{B\cap\ddot{C}},x_-\right)}\\ &= \frac{\mathbb{P}\left(X_{\ddot{i}}=1|x_{B\cap\ddot{C}},x_+\right) - \mathbb{P}\left(X_{\ddot{i}}=1|x_{B\cap\ddot{C}}\right)}{\mathbb{P}\left(X_{\ddot{i}}=1|x_{B\cap\ddot{C}},x_+\right) - \mathbb{P}\left(X_{\ddot{i}}=1|x_{B\cap\ddot{C}},x_-\right)}.\end{aligned}$$

and

- $\mathbb{P}\left(X_{\ddot{i}}=1|x_{B\cap\ddot{C}},x_-\right)$ is fed into $f_T(\cdot)$ for this lower bound

- $\mathbb{P}\left(X_{\ddot{i}} = 1 | x_{B\cap\ddot{C}}, x_+\right)$ is fed into $f_T(\cdot)$ for this lower bound

- $\mathbb{P}\left(X_{\ddot{i}} = 1 | x_{B\cap\ddot{C}}\right)$ is fed into $\hat{\mu}_T^{\ddot{i}}(B) \equiv \mathbb{E}_{[X_{B\cap\ddot{C}}]} f_T(\mathbb{P}\left(X_{\ddot{i}} = 1 | X_{B\cap\ddot{C}}\right))$, which is the corresponding upper bound

4.2 The single parent network

Recall the single parent network in Section 2.1 with N children. If the first m children are observed, the conditional probability of any another sibling $c > m$ is given by

$$\mathbb{P}\left(X_c = 1 | x_{\{1,\cdots,m\}}\right) = \mathbb{E}_{[X_P | x_{\{1,\cdots,m\}}]} \mathbb{P}\left(X_c = 1 | X_P\right),$$

where

$$\mathbb{P}\left(X_P | x_{\{1,\cdots,m\}}\right) = \frac{\mathbb{P}\left(X_P = 1\right) \mathbb{P}\left(x_{\{1,\cdots,m\}} | X_P = 1\right)}{\mathbb{E}_{[X_P]} \mathbb{P}\left(x_{\{1,\cdots,m\}} | X_P\right)}.$$

To calculate the exact value of a measure term $\mathbb{E}_{[X_B]} f_T(\mathbb{P}\left(X_{\ddot{i}} = 1 | X_B\right))$ for an observation set B, we have to consider the $2^{|B|}$ possible assignments to X_B. If we were to compare all possible observation sets of size m, we would have to consider 2^m assignments for each of the possible $\binom{N}{m}$ observation sets B. However, lets approximate the effect from the first k children whether they are included in B or not, and do exact calculations for the last $N - k$ variables $\ddot{C} = \{k+1, \cdots, c-1, c+1, \cdots, N\}$ to get a lower bound of the form (2). Now, x_- refers to $X_i = 0$ for all $i \le k$ and x_+ refers to $X_i = 1$ for all $i \le k$. In this case one can show that $w^{\ddot{i}}(x_{B\cap\ddot{C}}) = t \cdot \frac{\mathbb{P}\left(x_{B\cap\ddot{C}} | x_+\right)}{\mathbb{P}\left(x_{B\cap\ddot{C}}\right)}$ and equivalently $1 - w^{\ddot{i}}(x_{B\cap\ddot{C}}) = t \cdot \frac{\mathbb{P}\left(x_{B\cap\ddot{C}} | x_-\right)}{\mathbb{P}\left(x_{B\cap\ddot{C}}\right)}$, where $t = \frac{\mathbb{P}(P=1|x_+) - \mathbb{P}(P=1)}{\mathbb{P}(P=1|x_+) - \mathbb{P}(P=1|x_-)}$ does not depend on $x_{B\cap\ddot{C}}$. The reader is referred to Appendix A.1 for details. However, this means

$$\breve{\mu}_T^{\ddot{i}}(B) = t \cdot \mathbb{E}_{[X_{B\cap\ddot{C}} | x_-]} f_T(\mathbb{P}\left(X_{\ddot{i}} = 1 | x_-, X_{B\cap\ddot{C}}\right)) + (1-t) \cdot \mathbb{E}_{[X_{B\cap\ddot{C}} | x_+]} f_T(\mathbb{P}\left(X_{\ddot{i}} = 1 | x_+, X_{B\cap\ddot{C}}\right)).$$

In this special case our LB is easier to interpret, as it is defined as a sum of two parts where each part is conditioned on an extreme message and weighted according to the effect the opposite extreme message has on P.

5 Synthetic illustrating example: Simple two parent net

Assume $2N + 2$ binary variables (with value 0 or 1), related as in Figure 3, with probability distribution determined by

$$\begin{aligned} &\mathbb{P}\left(P_1 = 1\right) = p > 0, && \mathbb{P}\left(P_2 = 1 | P_1\right) = \rho \cdot P_1, \\ &\mathbb{P}\left(1_m = 1 | P_1\right) = p \cdot P_1, && \mathbb{P}\left(2_m = 1 | P_2\right) = p \cdot P_2. \end{aligned}$$

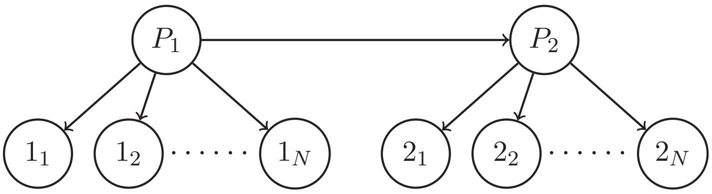

Figure 3. A BN with two sibling-parents with equal number of leaf-node children.

Assume the leaf nodes are the observable nodes and that the distribution is MTP$_2$. We can study the lower bound from Section 4 together with the upper bound from Section 3 for this network, and compare with the upper and lower bound from Section 3. We assume the variance measure, which is the measure discussed in Lilleborge and Eidsvik (2015). First, just look at values for the bounds and the exact measure value when B consists of one child of each of P_1 and P_2, i.e. $B = \{1_1, 2_1\}$. We calculate the effect from sibling observable nodes exactly, and approximate the effects of the others. First, the exact measure value

$$\begin{aligned}\frac{\mu_{Var}(\{1_1,2_1\})}{N-1} &= p^2(1+2\rho-\rho p)f_{Var}(p) + (1-p^2-\rho p^2+\rho p^3)f_{Var}\left(\frac{p-\rho p^2}{1+p-\rho p^2}\cdot p\right)\\ &\quad + p^2(1-\rho p)f_{Var}\left(\frac{\rho-\rho p}{1-\rho p}\cdot p\right) + (1-p^2-\rho p^2+\rho p^3)f_{Var}\left(\frac{\rho p-\rho p^2}{1+p-\rho p^2}\cdot p\right),\end{aligned}$$

then the upper bound

$$\frac{\hat{\mu}_{Var}(\{1_1,2_1\})}{N-1} = p^2(1+\rho)f_{Var}(p) + (1-p^2)f_{Var}\left(\frac{p}{1+p}\cdot p\right) + (1-\rho p^2)f_{Var}\left(\frac{\rho p-\rho p^2}{1-\rho p^2}\cdot p\right),$$

which is used together with the clairvoyant lower bound

$$\frac{\tilde{\mu}_{Var}(\{1_1,2_1\})}{N-1} = p(\rho+p)f_{Var}(p)+p(1-\rho p)f_{Var}\left(\frac{\rho-\rho p}{1-\rho p}\cdot p\right)+(1-p^2-\rho p+\rho p^2)f_{Var}\left(\frac{p-\rho p}{1+p-\rho p}\cdot p\right)$$

or the MTP$_2$ lower bound

$$\begin{aligned}\frac{\breve{\mu}_{Var}(\{1_1,2_1\})}{N-1} &= \left[p(\rho+p)-\rho p(1-p)^{N+1}\right]f_{Var}(p) + p\left(1-(1-p)^N\right)(1-\rho p)f_{Var}\left(\frac{\rho-\rho p}{1-\rho p}\cdot p\right)\\ &\quad + \left(1-p^2-\rho p+\rho p^2+\rho p(1-p)^{N+1}\right)f_{Var}\left(\frac{p-\rho p\left(1-(1-p)^N\right)}{1+p-\rho p\left(1-(1-p)^N\right)}\cdot p\right)\\ &\quad + \left(1-p+p(1-\rho p)(1-p)^N\right)f_{Var}\left(\frac{\rho p(1-p)^N}{1+p(1-\rho p)(1-p)^{N-1}}\cdot p\right).\end{aligned}$$

For the MTP$_2$ lower bound, each of the N children on a given side (left or right, respectively) gives an imperfect and independent indication on their parent (P_1 or P_2, respectively). The updated distribution of this parent is used for calculation of the bounds for the children on the other side (right or left, respectively). Compared to the clairvoyant bound, where one receives perfect information on the parent of the other side children, the MTP$_2$ lower bound receives weaker indications. Thus, $\tilde{\mu}_{Var}(B) \leq \breve{\mu}_{Var}(B)$ when $B \subseteq \{1_1, \cdots, 1_N, 2_1, \cdots, 2_N\}$. Further, observe that $\frac{\breve{\mu}_{Var}(B)}{N-1} \to \frac{\tilde{\mu}_{Var}(B)}{N-1}$ as $N \to \infty$. In fact, as $N \to \infty$, the indications on the other side parents in the MTP$_2$ lower bound are converging towards perfect information. That is, for N independent identically distributed boolean random variables $Y_1, \cdots, Y_N$ with success probability $p > 0$, $\mathbb{P}(\max\{Y_1, \cdots, Y_N\} = 1) \to 1$ as $N \to \infty$. Similarly for the two-parent net, $\mathbb{P}(\max\{1_1, \cdots, 1_N\} = P_1) \to 1$ and $\mathbb{P}(\max\{2_1, \cdots, 2_N\} = P_2) \to 1$ as $N \to \infty$.

The three bounds are plotted together with the true measure value in Figure 4 for different values of ρ. We see that for small ρ, e.g. when P_1 and P_2 are less correlated, the bounds are tight,

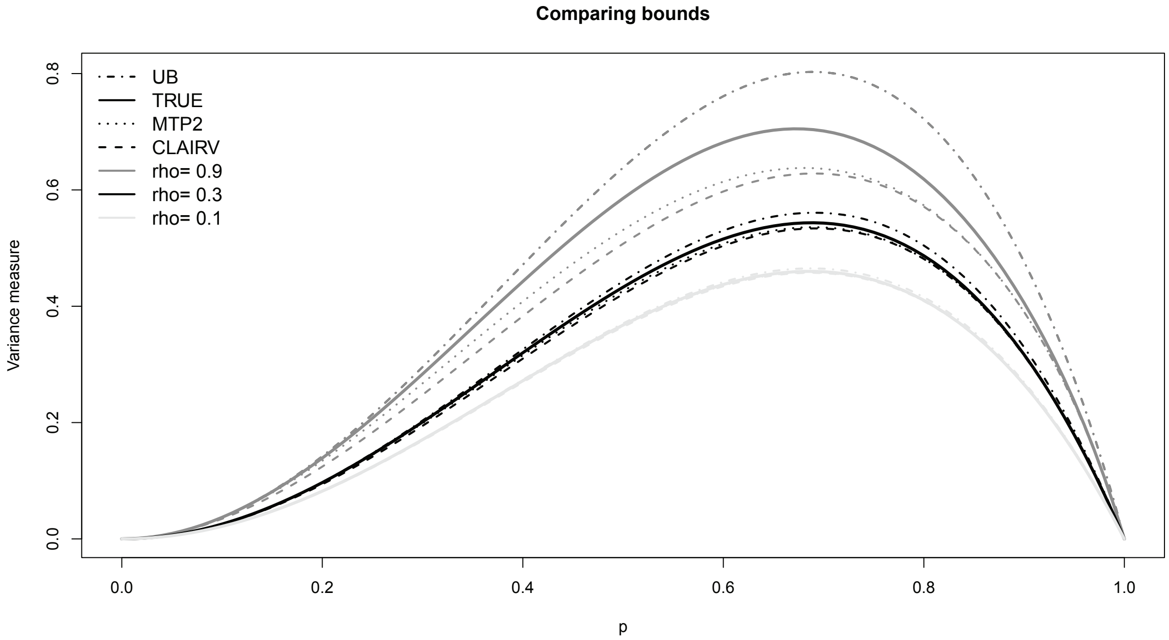

Figure 4. Comparing bounds for observation set $\{1_1, 2_1\}$ in the synthetic network in Figure 3 for $N = 3$. For each value of ρ (values and colors in legend), there is one curve for each of μ_T (True Value, solid line), $\hat{\mu}_T$ (Upper Bound, dashed-dotted line) $\tilde{\mu}_T$ (Clairvoyant Lower Bound, dashed line) and $\breve{\mu}_T$ (MTP$_2$ Lower Bound, dotted line) as a function of the success parameter p. The Clairvoyant bound is below the MTP$_2$ bound for all p (equal at end points).

while for ρ large, e.g. when P_1 and P_2 are highly correlated, the bounds are loose/conservative. The reason that the bounds are very loose in this case, is that the assumption for the upper bound (P_1 and P_2 independent) is totally off. Correspondingly for the Clairvoyant Lower Bound, the extra information (knowing the true value of P_i for the children of P_{2-i}) is too strong compared to the information from the actual observations. The MTP$_2$ Lower Bound give extra information, but less than the Clairvoyant Lower Bound, and provides a uniformly better lower bounds with similar behaviour. Also note that for a given ρ, the bounds are loosest around $p \sim 0.6$ to 0.8.

Recall that we are not necessarily looking for the uniformly tightest bound; we want our bound to be tight enough to ensure an optimal (or close-to optimal) observation set B_m of size m. The true measure values for the observation sets of two siblings are as follows,

$$\begin{aligned}\mu_{Var}(\{1_1, 1_2\}) = (N-2)p^2(2-p) \cdot f_{Var}(p) + (N-2)\left(1-p^2(2-p)\right) \cdot f_{Var}\left(\frac{p(1-p)}{1+p(1-p)} \cdot p\right) \\ + Np^2(2-p) \cdot f_{Var}(\rho p) + N\left(1-p^2(2-p)\right) \cdot f_{Var}\left(\frac{\rho p(1-p)}{1+p(1-p)} \cdot p\right)\end{aligned}$$

and

$$\begin{aligned}\mu_{Var}(\{2_1, 2_2\}) = (2N-2)\rho p^2(2-p) \cdot f_{Var}(p) + N\left(1-\rho p^2(2-p)\right) \cdot f_{Var}\left(\frac{p-\rho p^2(2-p)}{1-\rho p^2(2-p)} \cdot p\right) \\ + (N-2)\left(1-\rho p^2(2-p)\right) \cdot f_{Var}\left(\frac{\rho p(1-p)^2}{1-\rho p^2(2-p)} \cdot p\right),\end{aligned}$$

and the bounds are calculated similarly as for $\{1_1, 2_1\}$. That is, the upper bounds are

$$\hat{\mu}_{Var}(\{1_1, 1_2\}) = (N-2)p^2(1-p)f_{Var}(p) + (N-2)(1-p^2(2-p))f_{Var}\left(\frac{p(1-p)}{1+p-p^2} \cdot p\right) + Nf_{Var}\left(\rho p^2\right)$$

and

$$\hat{\mu}_{Var}(\{2_1, 2_2\}) = (N-2)\rho p^2(1-p)f_{Var}(p) + (N-2)(1-\rho p^2(2-p))f_{Var}\left(\frac{\rho p(1-p)^2}{1-\rho p^2(2-p)} \cdot p\right) + Nf_{Var}\left(p^2\right),$$

while the lower bounds are given by

$$\begin{aligned}\tilde{\mu}_{Var}(\{1_1, 1_2\}) =& (N-2)(p^2(2-p) + \rho p(1-p)^2)f_{Var}(p) + Npf_{Var}(\rho p) \\ &+ (N-2)(1-p^2(2-p) - \rho p(1-p)^2)f_{Var}\left(\frac{(1-\rho)(1-p)p^2}{1+p(1-p)(1-\rho)}\right)\end{aligned}$$

and

$$\begin{aligned}\tilde{\mu}_{Var}(\{2_1, 2_2\}) =& \left(N\rho p + (N-2)\rho p^2(2-p)\right) f_{Var}(p) + N(1-\rho p)f_{Var}\left(\frac{(1-\rho)p^2}{1-\rho p}\right) \\ &+ (N-2)(p - \rho p^2(2-p))f_{Var}\left(\frac{\rho(1-p)^2 p}{1-\rho p(2-p)}\right)\end{aligned}$$

for the Clairvoyant Lower Bound and

$$\begin{aligned}\breve{\mu}_{Var}(\{1_1,1_2\}) = (N-2)&\left(p^2(2-p) + \rho p(1-p)^2\left(1-(1-p)^N\right)\right) f_{Var}(p) + Np\left(1-(1-p)^N\right) f_{Var}(\rho p) \\ &+ (N-2)\left(1-p^2(2-p) - \rho p(1-p)^2\left(1-(1-p)^N\right)\right) f_{Var}\left(\frac{(1-p)\left(1-\rho+\rho(1-p)^N\right)p^2}{1+p(1-p)\left(1-\rho+\rho(1-p)^N\right)}\right) \\ &+ N\left(p(1-p)^N + 1 - p\right) f_{Var}\left(\frac{\rho(1-p)^N p^2}{p(1-p)^N + 1 - p}\right)\end{aligned}$$

and

$$\begin{aligned}\breve{\mu}_{Var}(\{2_1,2_2\}) = {} & \left(N\rho p\left(1-(1-p)^N\right) + (N-2)\rho p^2(2-p)\right) f_{Var}(p) \\ & + N(\rho p(1-p)^N + 1 - \rho p) f_{Var}\left(\frac{\rho p^2(1-p)^N + p^2(1-\rho)}{\rho p(1-p)^N + 1 - \rho p}\right) \\ & + (N-2)(p-\rho p^2(2-p))\left(1-(1-p)^N\right) f_{Var}\left(\frac{\rho(1-p)^2 p}{1-\rho p(2-p)}\right) \\ & + (N-2)\left(1-p+p(1-p)^N(1-\rho p(2-p))\right) f_{Var}\left(\frac{\rho(1-p)^{N+2}p^2}{1-p+p(1-p)^N(1-\rho p(2-p))}\right)\end{aligned}$$

for the MTP$_2$ Lower Bound.

Observe that the optimal $B \in \mathcal{B}_2$ of size 2 depends on N. Figure 5-Figure 8 show the bounds together with the true value for the different candidates when $N = 3$ for $\rho = 0.1$, 0.3 or 0.9, respectively. Recall that each candidate in $\mathcal{B}_2 = \{i_k, j_\ell \ : \ i,j \in \{1,2\},\ k,\ell \in \{1,\cdots,N\}\}$ is equivalent to either $\{1_1,2_1\}$, $\{1_1,1_2\}$ or $\{2_1,2_2\}$. Observe that for small p, e.g. when the observation nodes are not too good indications on their parents, its optimal to observe two children of P_1 in order to get a 'sufficiently good' indicator of this one parent. For larger p it is optimal to sample evidence on both parents, e.g. observe $1_1, 2_1$, since this gives information on both sides of the network. Note that the smallest p for which it is optimal to observe $\{1_1,2_1\}$ is larger for small ρ. Since the MTP$_2$ bound is uniformly tighter than the clairvoyant, there is an interval of p-values where the MTP$_2$ Lower Bound can separate out suboptimal candidates

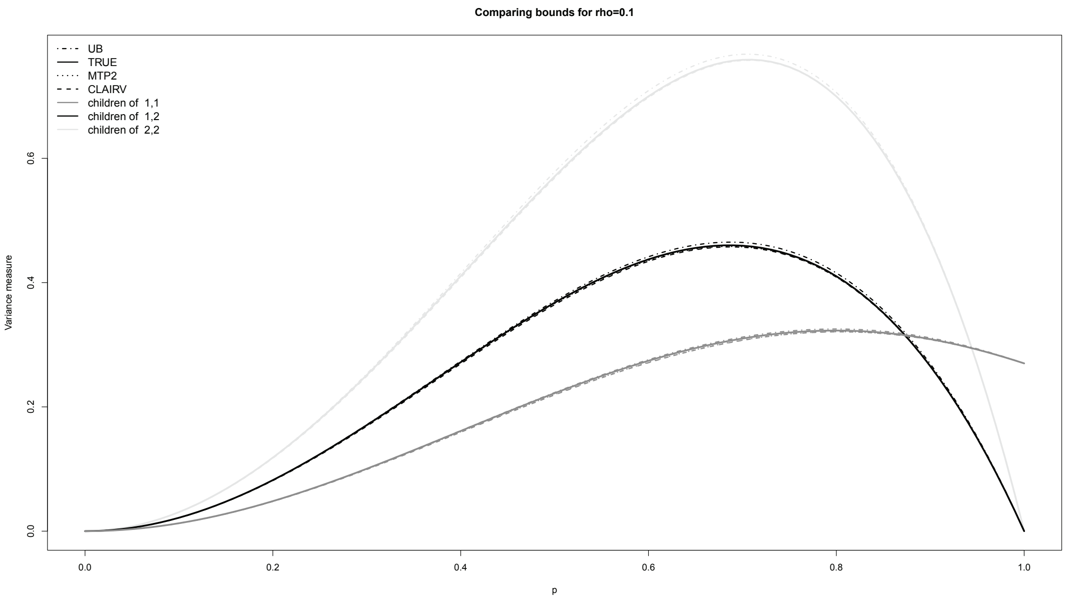

Figure 5. Comparing measure values and bounds for different observation sets in the synthetic network in Figure 3 for $\rho = 0.1$. There is one curve for each of $\mu_T(\{1_1,2_1\})$ (black curve), $\mu_T(\{1_1,1_2\})$ (dark gray curve) and $\mu_T(\{2_1,2_2\})$ (light gray curve) as a function of the success parameter p, together with the upper and lower bounds.

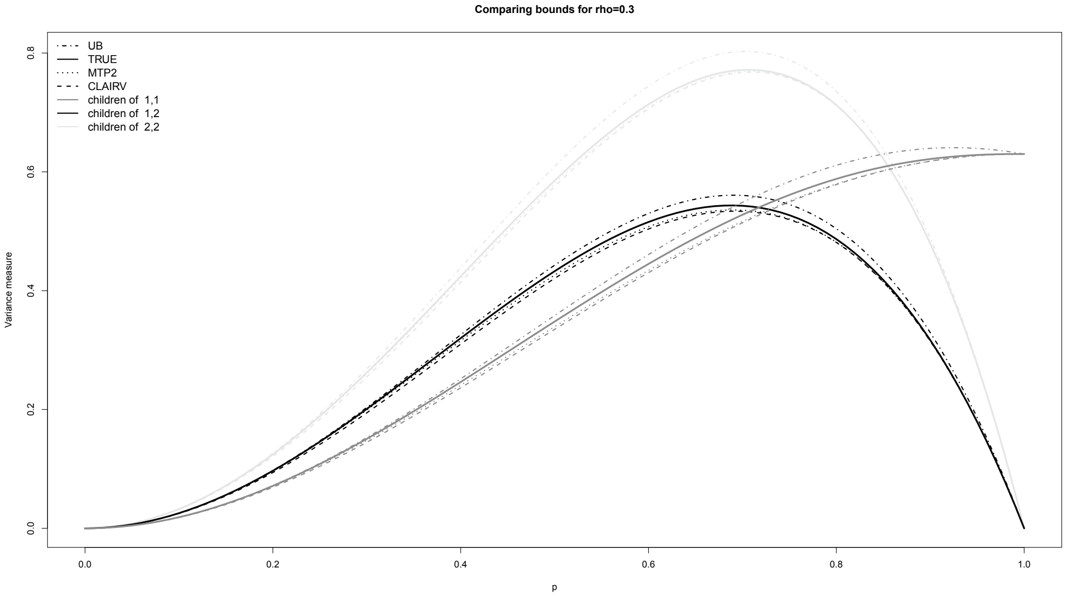

Figure 6. Comparing measure values and bounds for different observation sets in the synthetic network in Figure 3for $\rho = 0.3$. There is one curve for each of $\mu_T(\{1_1, 2_1\})$ (black curve), $\mu_T(\{1_1, 1_2\})$ (dark gray curve) and $\mu_T(\{2_1, 2_2\})$ (light gray curve) as a function of the success parameter p, together with the upper and lower bounds.

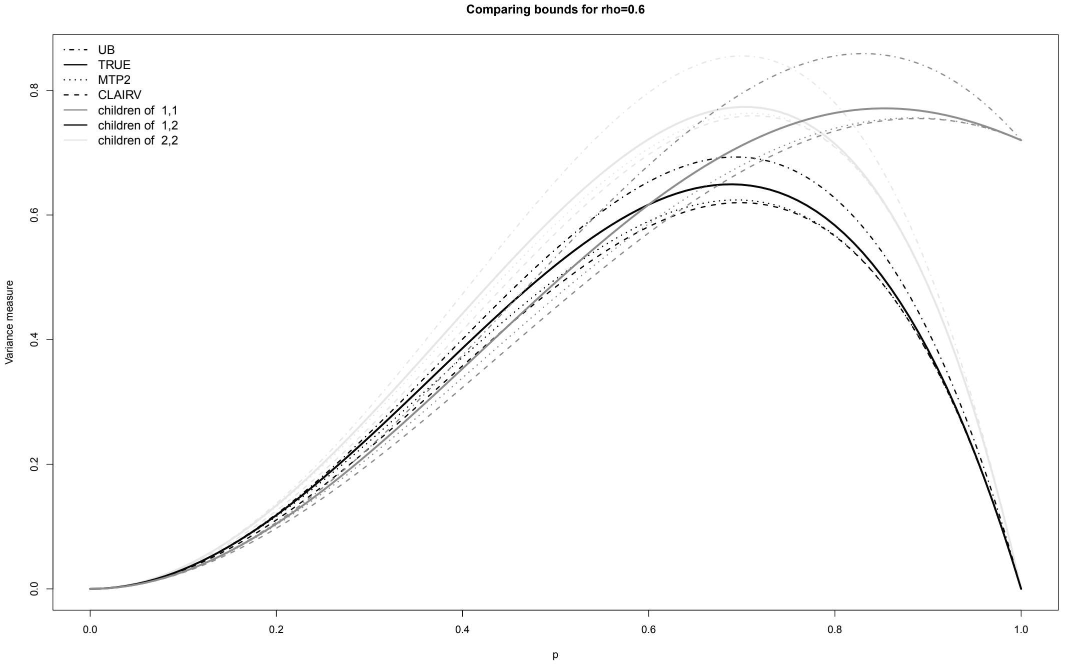

Figure 7. Comparing measure values and bounds for different observation sets in the synthetic network in Figure 3 for $\rho = 0.6$. There is one curve for each of $\mu_T(\{1_1, 2_1\})$ (black curve), $\mu_T(\{1_1, 1_2\})$ (dark gray curve) and $\mu_T(\{2_1, 2_2\})$ (light gray curve) as a function of the success parameter p, together with the upper and lower bounds.

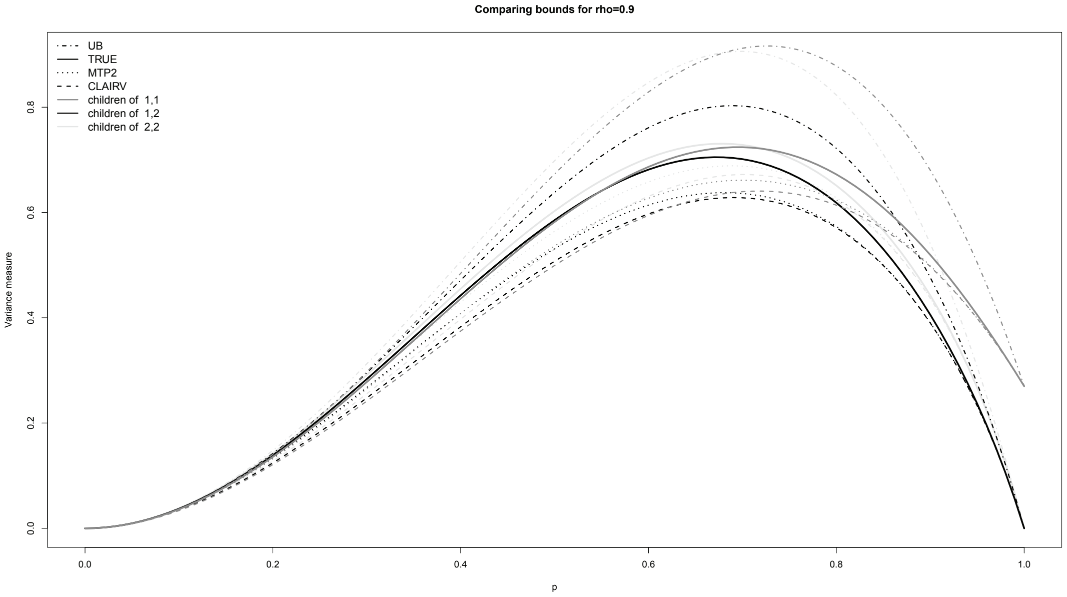

Figure 8. Comparing measure values and bounds for different observation sets in the synthetic network in Figure 3 for $\rho = 0.9$. There is one curve for each of $\mu_T(\{1_1, 2_1\})$ (black curve), $\mu_T(\{1_1, 1_2\})$ (dark gray curve) and $\mu_T(\{2_1, 2_2\})$ (light gray curve) as a function of the success parameter p, together with the upper and lower bounds.

	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.9$
Clairvoyant LB	$[0.871031, 0.876648]$	$[0.680558, 0.746506]$	$[0.030625, 0.971558]$
MTP$_2$ LB	$[0.871061, 0.876623]$	$[0.683610, 0.744892]$	$[0.109633, 0.971547]$

Table 1. Intervals of p-values for which the combination of upper and lower bound (Clairvoyant and MTP$_2$ Lower Bound, respectively) is indecisive of the optimal observation set. For smaller p-values, $\{1_1, 1_2\}$ is optimal, and for larger p-values $\{1_1, 2_1\}$ is optimal. The numbers are obtained numerically in an experiment where the bounds was compared for each p-value between 0 and 1 in increments of 10^{-6}.

(using the upper bound) while the Clairvoyant Lower Bound cannot, see Table 1. In general we see that this indecisive interval for the Clairvoyant Lower Bound is not much larger than for the MTP$_2$ Lower Bound. For small ρ (e.g. $\rho = 0.1$ in Table 1), both combinations of upper and lower bounds works well for most p-values. However, for large ρ (e.g. $\rho = 0.9$ in Table 1), both combinations are indecisive for most p-values, but the MTP$_2$ combination is successful for more than twice as many p-values.

When $\rho = 0.1$, with measure values and bounds visualized in Figure 5, the right part of the BN (Figure 3) is most likely "dead" as in all variables 0/failure. Thus, we need p large to have enough trust a single sample on the left to take the (small) chance of exploring a success on the right compared to the extra security in having a double sample in the more likely area. Also note that for large ρ (e.g. Figure 8, $\rho = 0.9$), all observations sets are close to optimal, since the indications on P_1 or P_2, respectively, are very similar for all children; because P_1 and P_2 are highly correlated. However, the bounds are too loose to give this indication: This is due to the same reason as discussed for Figure 4, unless p is very small or very large. For small ρ (see Figure 5 for $\rho = 0.1$ or Figure 6 for $\rho = 0.3$) the bounds are very good indicators for the optimal observation set. For ρ-values in between (see Figure 7, $\rho = 0.6$), the bounds are tight enough to either ensure the optimal observation set (for large p) or indicate that all observation sets are close to optimal (for small p).

6 Closing remarks

The lower bound constructed in Section 4 can be combined with the upper bound in Section 3 to optimize the subset selection problem as in Lilleborge and Eidsvik (2015). The set function optimized must be convex, as is the case for information measures as well as in many cases involving value of information. As this lower bound is targeted to MTP$_2$ distributions, they can utilize the MTP$_2$ properties to make an efficient bound. As in Lilleborge and Eidsvik (2015), the bounds can be applied to find the exact optimal solution by iteratively removing candidates as they are proved suboptimal by the bounds, or they can be used to speed up approximative search algorithms. The reader is referred to Lilleborge and Eidsvik (2015) for description and a discussion of such search algorithms.

Acknowledgements

This work is funded by Statistics for Innovation, $(\text{sfi})^2$, one of the Norwegian Centres for Research-based Innovation. The author also thanks the authors of Fallat et al. (2016) as well as the organizers (especially prof. Bala Rajaratnam) for inspirational sessions at the American Institute of Mathematics workshop “Positivity, Graphical Models, and the Modeling of Complex Multivariate Dependencies".

A Calculations

A.1 A special case of the MTP$_2$ bound: Updates for a JT clique with a single BN node as its unique separator

Consider a JT separator $\{P\}$ containing a single BN node P. We will study the probability updates for one of Ps BN-children C, which is included in a (unique) JT-neighbor of $\{P\}$. Lets for simplicity assume this JT-node represents $\{P, C\}$ and has no other neighbors. We will consider all possible messages from a given $\ddot{\mathcal{C}} \subset \mathrm{Ne}(\{P\})$ with $\{P, C\} \in \ddot{\mathcal{C}}$ and approximate with the extremes for the remaining JT neighbors of $\{P\}$. Let $\mathcal{C}$ be the collection of JT nodes in the direction of $\ddot{\mathcal{C}}$ from $\{P\}$, i.e. all nodes from which there is a path to $\{P\}$ going through a node in $\ddot{\mathcal{C}}$. Let $\ddot{L} = L \cap (\cup_{C \in \mathcal{C}} C)$ be the set of observable BN nodes represented in $\mathcal{C}$. We will accept the probability updates from observing nodes in $\ddot{L}$, and just consider the extremes x_- (i.e. $X_i = 0, i \in L \setminus \ddot{L}$) and x_+ (i.e. $X_i = 1, i \in L \setminus \ddot{L}$) for all other observable nodes. Also note that one can first calculate the effect of each extreme on the $\mathcal{C}$-subtree and then do all remaining calculations locally on the subtree.

We define $p_+ \equiv \mathbb{P}\left(P = 1 | x_+\right)$, $p_- \equiv \mathbb{P}\left(P = 1 | x_-\right)$ and $p = \mathbb{P}\left(P = 1\right)$, for convenient notation, and assume $p_+ > p_-$. (This will happen unless $L \setminus \ddot{L} \perp P$, i.e. the $\mathcal{C}$-subtree is actually independent of the remaining observable nodes in the original BN. In that case, the corresponding terms in the upper bound are the true values and there is no need for weights w.) Let $t = \frac{p_+ - p}{p_+ - p_-}$ such that t is the unique solution of $p = t \cdot p_- + (1 - t) \cdot p_+$.

Given an assignment $X_B = x_B,\ x_B \in \chi_B$, it has a restricted assignment $X_{B \cap \ddot{L}} = x_{B \cap \ddot{L}}$ with corresponding $q_1 = \mathbb{P}\left(x_{B \cap \ddot{L}} | P = 1\right)$ and $q_0 = \mathbb{P}\left(x_{B \cap \ddot{L}} | P = 0\right)$. Now,

$$\mathbb{P}\left(P = 1 | x_\pm, x_{B \cap \ddot{L}}\right) = \frac{\mathbb{P}\left(P = 1, x_{B \cap \ddot{L}} | x_\pm\right)}{\mathbb{P}\left(x_{B \cap \ddot{L}} | x_\pm\right)} = \frac{q_1 \cdot p_\pm}{\mathbb{P}\left(x_{B \cap \ddot{L}} | x_\pm\right)}, \qquad \mathbb{P}\left(x_{B \cap \ddot{L}} | x_\pm\right) = q_1 \cdot p_\pm + q_0 \cdot (1 - p_\pm)$$

$$\mathbb{P}\left(P = 1 | x_{B \cap \ddot{L}}\right) = \frac{p \cdot q_1}{\mathbb{P}\left(x_{B \cap \ddot{L}}\right)}, \qquad \mathbb{P}\left(x_{B \cap \ddot{L}}\right) = q_1 \cdot p + q_0 \cdot (1 - p)$$

These equations let us specify $w^{\ddot{i}}(x_{B\cap\ddot{L}})$ further, as

$$\begin{aligned}
w^{\ddot{i}}(x_{B\cap\ddot{L}}) \;=\; \mathbb{E}_{[X_{B\setminus\ddot{L}}|X_{B\cap\ddot{L}}]}t_{X_B} &= \frac{\mathbb{P}\left(P=1|x_{B\cap\ddot{L}},x_+\right)-\mathbb{P}\left(P=1|x_{B\cap\ddot{L}}\right)}{\mathbb{P}\left(P=1|x_{B\cap\ddot{L}},x_+\right)-\mathbb{P}\left(P=1|x_{B\cap\ddot{L}},x_-\right)}\\
&= \left(\frac{p_+\cdot q_1}{\mathbb{P}\left(x_{B\cap\ddot{L}}|x_+\right)}-\frac{p\cdot q_1}{\mathbb{P}\left(x_{B\cap\ddot{L}}\right)}\right)\Big/\left(\frac{p_+\cdot q_1}{\mathbb{P}\left(x_{B\cap\ddot{L}}|x_+\right)}-\frac{p_-\cdot q_1}{\mathbb{P}\left(x_{B\cap\ddot{L}}|x_-\right)}\right)\\
&= \frac{p_+\cdot\mathbb{P}\left(x_{B\cap\ddot{L}}\right)-p\cdot\mathbb{P}\left(x_{B\cap\ddot{L}}|x_+\right)}{p_+\cdot\mathbb{P}\left(x_{B\cap\ddot{L}}|x_-\right)-p_-\cdot\mathbb{P}\left(x_{B\cap\ddot{L}}|x_+\right)}\cdot\frac{\mathbb{P}\left(x_{B\cap\ddot{L}}|x_-\right)\cdot\mathbb{P}\left(x_{B\cap\ddot{L}}|x_+\right)}{\mathbb{P}\left(x_{B\cap\ddot{L}}|x_+\right)\cdot\mathbb{P}\left(x_{B\cap\ddot{L}}\right)}\\
&= \frac{p_+\cdot(p\cdot q_1+(1-p)\cdot q_0)-p\cdot(p_+\cdot q_1+(1-p_+)\cdot q_0)}{p_+\cdot(p_-\cdot q_1+(1-p_-)\cdot q_0)-p_-\cdot(p_+\cdot q_1+(1-p_+)\cdot q_0)}\cdot\frac{\mathbb{P}\left(x_{B\cap\ddot{L}}|x_-\right)}{\mathbb{P}\left(x_{B\cap\ddot{L}}\right)}\\
&= \frac{q_0\,(p_+-p)}{q_0\,(p_+-p_-)}\cdot\frac{\mathbb{P}\left(x_{B\cap\ddot{L}}|x_-\right)}{\mathbb{P}\left(x_{B\cap\ddot{L}}\right)}\\
&= t\cdot\frac{\mathbb{P}\left(x_{B\cap\ddot{L}}|x_-\right)}{\mathbb{P}\left(x_{B\cap\ddot{L}}\right)},
\end{aligned}$$

and correspondingly,

$$\begin{aligned}
1-w^{\ddot{i}}(x_{B\cap\ddot{L}}) &= \frac{\mathbb{P}\left(x_{B\cap\ddot{L}}\right)-t\cdot\mathbb{P}\left(x_{B\cap\ddot{L}}|x_-\right)}{\mathbb{P}\left(x_{B\cap\ddot{L}}\right)}\\
&= \frac{(p\cdot q_1+(1-p)\cdot q_0)-t\cdot(p_-\cdot q_1+(1-p_-)\cdot q_0)}{\mathbb{P}\left(x_{B\cap\ddot{L}}\right)}\\
&= \frac{(p_-\cdot t+p_+\cdot(1-t))\cdot q_1+(1-p_-\cdot t-p_+\cdot(1-t))\cdot q_0-t\cdot(p_-\cdot q_1+(1-p_-)\cdot q_0)}{\mathbb{P}\left(x_{B\cap\ddot{L}}\right)}\\
&= (1-t)\frac{\cdot(p_+\cdot q_1+(1-p_+)\cdot q_0)}{\mathbb{P}\left(x_{B\cap\ddot{L}}\right)}\\
&= (1-t)\frac{\cdot\mathbb{P}\left(x_{B\cap\ddot{L}}|x_+\right)}{\mathbb{P}\left(x_{B\cap\ddot{L}}\right)}.
\end{aligned}$$

Finally,

$$\begin{aligned}
\breve{\mu}_T^{\ddot{i}}(B) &\equiv \mathbb{E}_{[X_{B\cap\ddot{L}}]}\left[w(X_{B\cap\ddot{L}})\cdot f_T(\mathbb{P}\left(X_{\ddot{i}}=1|x_-,X_{B\cap\ddot{L}}\right))+\left(1-w(X_{B\cap\ddot{L}})\cdot f_T(\mathbb{P}\left(X_{\ddot{i}}=1|x_+,X_{B\cap\ddot{L}}\right))\right)\right]\\
&= t\cdot\mathbb{E}_{[X_{B\cap\ddot{L}}|x_-]}f_T(\mathbb{P}\left(X_{\ddot{i}}=1|x_-,X_{B\cap\ddot{L}}\right))+(1-t)\cdot\mathbb{E}_{[X_{B\cap\ddot{L}}|x_+]}f_T(\mathbb{P}\left(X_{\ddot{i}}=1|x_+,X_{B\cap\ddot{L}}\right)),
\end{aligned}$$

i.e. it is defined as a sum of two parts, each part conditioned on an ”extreme” message and weighted by t. This makes the lower bound easier to interpret in this case.

References

Almond, R. and Kong, A. (1991). Optimality issues in constructing a markov tree from graphical models. Technical report, Department of Statistics, Harvard University.

Bonneau, M., Gaba, S., Peyrard, N., and Sabbadin, R. (2014). Reinforcement learning-based design of sampling policies under cost constraints in markov random fields: Application to weed map reconstruction. *Computational Statistics & Data Analysis*, 72:30–44.

Brown, D. and Smith, J. (2013). Optimal Sequential Exploration: Bandits, Clairvoyants, and Wildcats. *Operations Research*, 61(3):644–665.

Fallat, S., Lauritzen, S., Sadeghi, K., Uhler, C., Wermuth, N., and Zwiernik, P. (2016). Total positivity in markov structures. Technical report, arXiv.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local Computation with Probabilities on Graphical Structures and their Application to Expert Systems (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 50(2):157–224.

Lilleborge, M. and Eidsvik, J. (2015). Efficient designs for bayesian networks with sub-tree bounds. *Statistics and Computing*, pages 1–18. Available from: `http://dx.doi.org/10.1007/s11222-015-9623-0`.

Lilleborge, M., Hauge, R., and Eidsvik, J. (2015). Information gathering in bayesian networks applied to petroleum prospecting. *Mathematical Geosciences*, 48(3):233–257. Available from: `http://dx.doi.org/10.1007/s11004-015-9616-8`.

Martinelli, G. and Eidsvik, J. (2014). Dynamic Exploration Designs for Graphical Models using Clustering with Applications to Petroleum Exploration . *Knowledge-Based Systems*, 58:113–126.

Peyrard, N., Sabbadin, R., Spring, D., Brook, B., and Mac Nally, R. (2013). Model-based adaptive spatial sampling for occurrence map construction. *Statistics and Computing*, 23(1):29–42.

IV