Most Frugal Explanations in Bayesian Networks

Johan Kwisthout*

Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, The Netherlands

Abstract

Inferring the most probable explanation to a set of variables, given a partial observation of the remaining variables, is one of the canonical computational problems in Bayesian networks, with widespread applications in AI and beyond. This problem, known as MAP, is computationally intractable (NP-hard) and remains so even when only an approximate solution is sought. We propose a heuristic formulation of the MAP problem, denoted as Inference to the Most Frugal Explanation (MFE), based on the observation that many intermediate variables (that are neither observed nor to be explained) are irrelevant with respect to the outcome of the explanatory process. An explanation based on few samples (often even a singleton sample) from these irrelevant variables is typically almost as good as an explanation based on (the computationally costly) marginalization over these variables. We show that while MFE is computationally intractable in general (as is MAP), it can be tractably approximated under plausible situational constraints, and its inferences are fairly robust with respect to which intermediate variables are considered to be relevant.

Keywords: Bayesian Abduction, Parameterized Complexity, Approximation, Heuristics, Computational Complexity

1. Introduction

Abduction or inference to the best explanation refers to the process of finding a suitable explanation (the *explanans*) of observed data or phenom-

*Corresponding author: Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, PO Box 9104, 6500HE Nijmegen, The Netherlands; e-mail j.kwisthout@donders.ru.nl; telephone +31 (0) 24 3616288

ena (the *explananda*). In the last decades, Bayesian notions of abduction have emerged due to the widespread popularity of Bayesian or probabilistic techniques for representing and reasoning with knowledge [5, 26, 30, 47, 52]. They are used in decision support systems in a wide range of problem domains [e.g., 7, 11, 21, 23, 32, 45, 64] and as computational models of economic, social, or cognitive processes [10, 25, 33, 48, 58, 60]. The natural interpretation of 'best' in such models is 'most probable': the explanation that is the most probable one given the evidence, i.e., that has maximum posterior probability, is seen as the hypothesis that best explains the available evidence; this explanation is traditionally referred to as the MAP explanation and the computational problem of inferring this explanation as the MAP problem.[1]

However, computing or even approximating the MAP explanation is computationally costly (i.e., NP-hard), especially when there are many intermediate (neither observed nor to be explained) variables that may influence the explanation [1, 4, 51, 56]. To compute the posterior probability distribution of the explanation variables, all these intermediate variables need to be marginalized over. One way of dealing with this intractability might be by assuming modularity of knowledge representations, i.e., by assuming that these representations are informationally encapsulated and do not have access to background knowledge. However, this is problematic as we cannot know beforehand which elements of background knowledge or observations may be relevant for determining the best explanation [17, 19].

Fortunately, even when a full Bayesian computation may not be feasible in large networks, we need not constrain inferences only to small or disconnected knowledge structures. It is known that in general the posterior probability distribution of a (discrete) Bayesian network is skewed, i.e., a few joint value assignments cover most of the probability space [13], and that typically only few of the variables in a network are relevant for a par-

[1]Other relationships have been proposed that compete in providing 'sufficiently rational' relations between observed phenomena and their explanation that can be used to describe *why* we judge one explanation to be preferred over another [28, 44]. Examples include *maximum likelihood* [29], which does not take the prior probabilities of the hypotheses into account, the *conservative Bayesian* approach [6], *generalized Bayes* factor [66], and various Bayesian formalisms of *coherence theory* [5, 15, 26, 49]. While the posterior probability of such explanations is not the deciding criterion to prefer one explanation over another, it is typically so that explanations we consider to be good for other reasons also have a high posterior probability compared to alternative explanations [27, 44].

ticular inference query [14]. We propose to utilize this property of Bayesian networks in order to make tractable (approximate) inferences to the best explanation over large and unencapsulated knowledge structures. We introduce a heuristic formulation of MAP, denoted as Inference to the Most Frugal Explanation (MFE), that is explicitly designed to reflect that only few intermediate variables are typically relevant in real-world situations. In this formulation we partition the set of intermediate variables in the network into a set of 'relevant' intermediate variables that are marginalized over, and a set of 'irrelevant' intermediate variables that we sample from in order to estimate an explanation.

Note that in the MFE formalism there is marginalization over *some* of the intermediate variables (the variables that are considered to be relevant), but not over those intermediate variables that are not considered to be relevant. Thus, MFE can be seen as a 'compromise' between computing the explanation with maximum posterior probability, where one marginalizes over all intermediate variables, and the previously proposed Most Simple Explanation (MSE) formalism [35] where there is no marginalization at all, i.e., all intermediate variables are seen as irrelevant. We want to emphasize that the notions 'relevant' and 'irrelevant' in the problem definition denote *subjective* partitions of the intermediate variables; we will revisit this issue in Section 3.1.

We claim that this heuristic formalism of the MAP problem exhibits the following desirable properties:

1. The knowledge structures are *isotropic*, i.e., they are such that, potentially, everything can be relevant to the outcome of an inference process. They are also *Quinean*: candidate explanations are sensitive to the entire belief system [17, 18].

2. The inferences are provably computationally tractable (either to compute exactly or to approximate) under realistic assumptions with respect to situational constraints [43, 53].

3. The resulting explanations have an optimal or close-to-optimal posterior probability in many cases, i.e., MFE actually 'tracks truth' in the terms of Glass [28].

In the remainder of this paper, we will discuss some needed preliminaries in Section 2. In Section 3 we discuss MFE in more detail. We give a more formal definition, including a formal definition of relevance in the context of Bayesian networks, and show how MFE can be tractably approximated under realistic assumptions despite computational intractability of the problem in general. In Section 4 we show that MFE typically gives an explanation that has a close-to-optimal posterior probability, even if only a subset of the relevant variables is considered. We discuss how MFE performs under various scenarios (e.g., when there are few or many relevant variables, when there are many hypotheses that are almost equally likely, or when there is a misalignment between the *actual* relevant variables and the variables that are mistakenly presumed to be relevant). We conclude our paper in Section 5.

2. Preliminaries

In this section we will introduce some preliminaries from Bayesian networks, in particular the MAP problem as standard formalization of Bayesian abduction. We will discuss the ALARM network which we will use as a running example throughout this paper. Lastly, we introduce some needed concepts from complexity theory, in particular the complexity class PP, oracles, and fixed parameter tractability.

2.1. Bayesian networks and Bayesian abduction

A Bayesian or probabilistic network $\mathcal{B}$ is a graphical structure that models a set of stochastic variables, the conditional independences among these variables, and a joint probability distribution over these variables [52]. $\mathcal{B}$ includes a directed acyclic graph $\mathbf{G}_{\mathcal{B}} = (\mathbf{V}, \mathbf{A})$, modeling the variables and conditional independences in the network, and a set of parameter probabilities Pr in the form of conditional probability tables (CPTs), capturing the strengths of the relationships between the variables. The network models a joint probability distribution $\Pr(\mathbf{V}) = \prod_{i=1}^{n} \Pr(V_i \mid \pi(V_i))$ over its variables, where $\pi(V_i)$ denotes the parents of V_i in $\mathbf{G}_{\mathcal{B}}$. We will use upper case letters to denote individual nodes in the network, upper case bold letters to denote sets of nodes, lower case letters to denote value assignments to nodes, and lower case bold letters to denote joint value assignments to sets of nodes. We will sometimes write $\Pr(\mathbf{x} \mid \mathbf{y})$ as a shorthand for $\Pr(\mathbf{X} = \mathbf{x} \mid \mathbf{Y} = \mathbf{y})$ if no ambiguity can occur.

In a Bayesian abduction task there are three types of variables: the *evidence* variables, the *explanation* variables, and a set of variables called *inter-*

mediate variables that are neither evidence nor explanation variables. The evidence variables are instantiated, i.e., have been assigned a value; the joint value assignment constitutes the explananda (what is to be explained, viz., the observations, data, or evidence). The explanation variables together form the *hypothesis space*: a set of possible explanations for the observations; a particular joint value assignment to these variables constitutes an explanans (the actual explanation of the observations). When determining what is the *best* explanation, typically we also need to consider other variables that are not directly observed, nor are to be explained: the intermediate variables. By convention, we will use $\mathbf{E}$, $\mathbf{H}$, and $\mathbf{I}$, to denote the sets of evidence variables, explanation variables, and intermediate variables, respectively. We will use $\mathbf{e}$ to denote the evidence, viz., the (observed) joint value assignment to the evidence variables.

The problem of inferring the *most probable* explanation, i.e., the joint value assignment for the explanation set that has maximum posterior probability given the evidence, is defined as MAP, or also Partial MAP or Marginal MAP to emphasize that the probability of any such joint value assignment is computed by marginalization over the intermediate variables. MAP is formally defined as follows.

Maximum A Posteriori Probability (MAP)
Instance: A Bayesian network $\mathcal{B} = (\mathbf{G}_{\mathcal{B}}, \Pr)$, where $\mathbf{V}$ is partitioned into evidence variables $\mathbf{E}$ with joint value assignment $\mathbf{e}$, explanation variables $\mathbf{H}$, and intermediate variables $\mathbf{I}$.
Output: The joint value assignment $\mathbf{h}$ to the nodes in $\mathbf{H}$ that has maximum posterior probability given the evidence $\mathbf{e}$.

2.2. The ALARM network

The ALARM network (Figure 1) will be used throughout this paper as a running example. This network is constructed as a part of the ALARM monitoring system, providing users with text messages denoting possible problems in anesthesia monitoring [2]. It consists of thirty-seven discrete random variables. Eight of these variables are designed as diagnostic variables that are to be explained, indicating problems like pulmonary embolism or a kinked tube; another sixteen variables indicate measurable or observable findings. The remaining thirteen variables are intermediate variables, i.e., they are neither diagnostic variables, nor can be observed (in principle or in practice). Apart from its practical use in the system described above, the ALARM network is one of the most prominent benchmark networks in the Bayesian network community.[2]

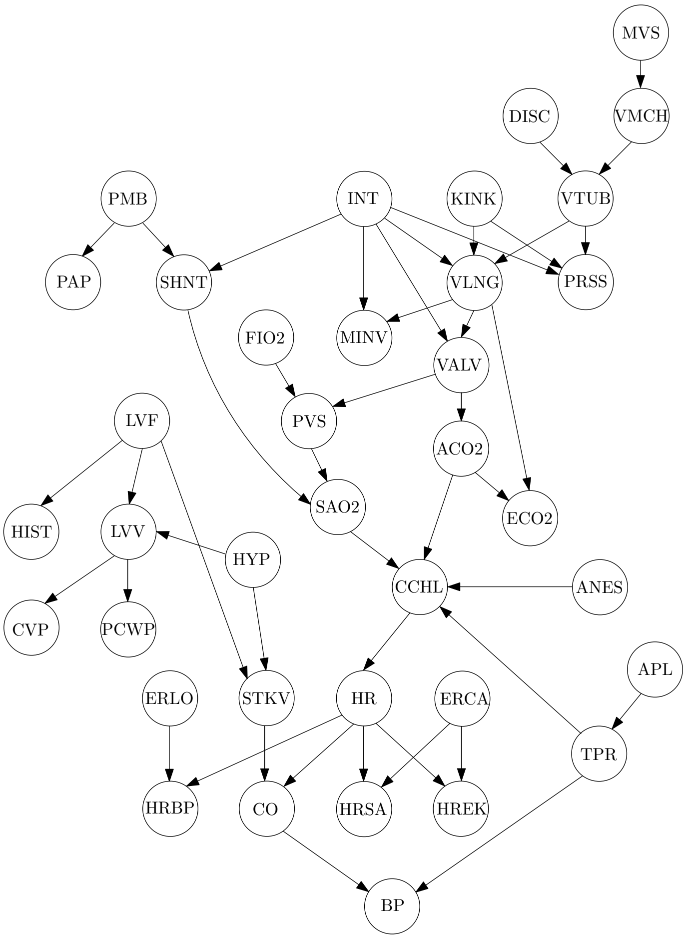

Figure 1: The ALARM network [2]

As an example, consider that a high breathing pressure was detected (PRSS = high) and that minute ventilation was low (MINV = low); all other observable variables take their default (i.e., non-alarming) value. From these findings a probability of 0.92 for the diagnosis 'kinked tube' (KINK =

[2] See, e.g., `http://www.cs.huji.ac.il/site/labs/compbio/Repository/`

true) can be computed. Likewise, we can compute that the most probable joint explanation for the diagnostic variables, given that PCWP (pulmonary capillary wedge pressure) and BP (blood pressure) are high, is that HYP = true (hypovolemia, viz., loss of blood volume) and all other diagnostic variables are negative. This joint value assignment has probability 0.58. The second-best explanation (all diagnostic variables are negative, despite the two alarming conditions) has probability 0.11.

2.3. Complexity theory

In the remainder, we assume that the reader is familiar with basic concepts of computational complexity theory, such as Turing Machines, the complexity classes P and NP, and intractability proofs. For more background we refer to classical textbooks like [22] and [50]. In addition to these basic concepts we will introduce concepts that are in particular relevant to Bayesian computations, in particular Probabilistic Turing Machines, Oracle Turing Machines, the complexity class PP and the Counting Hierarchy; the interested reader will find more background in [34] or [8]. Finally, we will briefly, and somewhat informally, introduce parameterized complexity theory. A more thorough introduction can be found in [12] or [16].

A Probabilistic Turing Machine (PTM) augments the more traditional Non-deterministic Turing Machine (NTM) with a probability distribution associated with each state transition. Without loss of generality we may assume that state transitions are binary and that the probability distribution at each transition is uniform. A PTM accepts a language L if the probability of ending in an accepting state when given some input x is strictly larger than $1/2$ if and only if $x \in L$. Given uniformly distributed binary state transitions this is exactly the case if the majority of computation paths accepts. The complexity class PP is defined as the class of languages accepted by some PTM in polynomial time. Observe that $\mathsf{NP} \subseteq \mathsf{PP}$; the inclusion is thought to be strict. PP contains complete problems, the canonical one being MAJSAT: given a Boolean formula ϕ, does the majority of truth assignments to the variables satisfy it?

An Oracle Turing Machine (OTM) is a Turing Machine enhanced with a so-called *oracle tape* and an oracle O for deciding membership queries for a particular language L_O. Apart from its usual operations, the OTM can write a string y on the oracle tape and 'summon the oracle'. In the next state, the OTM will have either replaced the string with 1 if $y \in L_O$, or 0 if $y \notin L_O$. The oracle can thus be seen as a 'black box' that answers membership queries in constant time. Note that both accepting and rejecting answers of the oracle can be used. Various complexity classes are defined using oracles; for example, the class $\mathsf{NP}^{\mathsf{PP}}$ includes exactly those languages that can be decided on an NTM with an oracle for PP-complete languages. Using the class PP and hierarchies of oracles the *Counting Hierarchy* [61] can be defined as a generalization of the Polynomial Hierarchy [59], including classes as $\mathsf{NP}^{\mathsf{PP}}$, $\mathsf{P}^{\mathsf{NP}^{\mathsf{PP}}}$, or $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$. Canonical complete problems for such classes include various SATISFIABILITY variants, using the quantifiers $\forall$, $\exists$, and MAJ to bind subsets of variables [61, 63].

Sometimes problems are intractable (i.e., NP-hard) in general, but become tractable if some *parameters* of the problem can be assumed to be small. Informally, a problem is called fixed-parameter tractable for a parameter k (or a set of parameters $\{k_1, \ldots, k_m\}$) if it can be solved in time, exponential (or even worse) *only* in k and polynomial in the input size $|x|$. In practice, this means that problem instances can be solved efficiently, even when the problem is NP-hard in general, if k is known to be small. If an NP-hard problem Π is fixed-parameter tractable for a particular parameter set k then k is denoted a *source of complexity* [53] of Π: bounding k renders the problem tractable, whereas leaving k unbounded ensures intractability under usual complexity-theoretic assumptions like $\mathsf{P} \neq \mathsf{NP}$. On the other hand, if Π remains NP-hard independent of the value of parameter k, then Π is para-NP-hard with respect to k: bounding k does not render the problem tractable. The notion of fixed-parameter tractability can be extended to deal with *rational*, rather than integer, parameters [36]. Informally, if a problem is fixed-rational tractable for a (rational) parameter k, then the problem can be solved tractably if k is close to 0 (or, depending on the definition, to 1). For readability, we will liberally mix integer and rational parameters in the remainder.

3. Most Frugal Explanations

In real-world applications there are many intermediate variables that are neither observed nor to be explained, yet may influence the explanation. Some of these variables can considerably affect the outcome of the abduction process. Most of these variables, however, are irrelevant as they are not expected to influence the outcome of the abduction process in all but maybe the very rarest of cases [14]. To compute the most probable explanation of the evidence, however, one needs to marginalize over all these variables, that is, take their prior or conditional probability distribution into account. This seems like a waste of computing resources in cases where we might as well have assigned an arbitrary value to these variables and still arrive at the same explanation.

One way of ensuring tractability of inference may be by 'weeding out' the irrelevant aspects in the knowledge structure prior to inference, reducing the network to a simplified version. For example, one might try to identify intermediate variables in the network that are conditionally independent of the explanation variables, given the evidence. While this can be done tractably in principle [24], it may still leave us with many variables that are conditionally dependent, yet do not influence the most probable explanation of the evidence. These variables are still in a sense redundant for finding explanations, as illustrated in the following example.

Example 1. Consider in the ALARM network the observations that PCWP and BP are high and the other observable variables take their non-alarming states. The actual value of ACO2 does not influence the most probable value of the observable variables in the network, i.e., $\mathrm{argmax}_{\mathbf{h}}\Pr(\mathbf{h}, \mathbf{e}, \mathbf{i}, \mathrm{ACO2} = \mathit{high}) = \mathrm{argmax}_{\mathbf{h}}\Pr(\mathbf{h}, \mathbf{e}, \mathbf{i}, \mathrm{ACO2} = \mathit{mid}) = \mathrm{argmax}_{\mathbf{h}}\Pr(\mathbf{h}, \mathbf{e}, \mathbf{i}, \mathrm{ACO2} = \mathit{low})$ for every joint value assignment $\mathbf{i}$ to the intermediate variables other than ACO2. However, ACO2 is not conditionally independent of (e.g.) KINK given the observed evidence variables.

An alternative to only selecting those intermediate variables that are conditionally dependent on the explanation variables is to apply a stronger criterion for relevance, e.g., selecting only those variables whose value may potentially change the most probable explanation. However, finding these variables itself would require potentially intractable computations as we will illustrate in Section 3.1 and formally prove in the Appendix. Furthermore, we might want to even constrain the number of variables to select even more by demanding not only that their value *might* change the most probable explanation (e.g., in some extraordinary combination of values for the other variables), but in fact actually *does* change the most probable explanation in a non-trivial number of situations. In addition, it is preferable to have a means of trading off the quality of a solution and the time needed to obtain a solution.

Example 2 (Adapted from [35])**.** Mr. Jones typically comes to work by train. Today Mr. Jones is late while he has been seen to leave his house at the usual time. One explanation can be that the train is delayed. However, it might also be the case that Mr. Jones was the unlucky individual who walked through 11th Street at 8.03 AM and was shot during an armed bank robbery, while mistakenly taken for a police constable. When trying to explain why Mr. Jones is not at his desk on 8.30 AM, there are a number of variables we might take into account, for example whether he has to change trains. A whole lot of variables are typically not taken into account because they are normally not relevant in most of the cases, for example the color of Mr. Jones's coat, or whether walked on the left or right pavement in 11th Street. Only in the awkward coincidence that Mr. Jones was in the wrong place at the wrong time they become relevant to explain why he is not at work.

Our approach is not to reduce the network to only include those intermediate variables we consider to be relevant and do inference on the resulting pruned network. In contrast, we propose that (the computationally costly) marginalization is done only on a subset of the intermediate variables (the variables that are considered to be relevant), and that a sampling strategy is used for the remaining intermediate variables that are not considered to be relevant. Such a sampling strategy may be very simple ('decide using a singleton sample') or more complex ('compute the best explanation on N samples and take a majority vote'). This allows for a trade-off between time to compute a solution and the quality of the result obtained, by having both a degree of freedom on which variables to include in the set of relevant intermediate variables and a degree of freedom on how many samples to take on the remaining intermediate variables. In Section 4 we will discuss the effects of such choices using computer simulations on random networks.

We now formally define the Most Frugal Explanation problem as follows[3]:

Most Frugal Explanation (MFE)
Instance: A Bayesian network $\mathcal{B}$, partitioned into a set of observed evidence variables $\mathbf{E}$, a set of explanation variables $\mathbf{H}$, a set of 'relevant'

[3]To improve readability, this formulation does not explicate how to deal with the following borderline cases: a) for any given joint value assignment to the irrelevant intermediate variables, multiple hypotheses have the same posterior probability; and b) multiple hypotheses are most probable for the same maximum number of (possibly distinct) hypotheses. The implementation of the algorithm described in Section 3.3 dealt with both these borderline cases by randomly selecting one of the competing hypotheses in case of a tie.

intermediate variables $\mathbf{I}^+$ that are marginalized over, and a set of 'irrelevant' intermediate variables $\mathbf{I}^-$ that are not marginalized over.
Output: The joint value assignment to the variables in the explanation set that is most probable for the maximum number of joint value assignments to the irrelevant intermediate variables.

The approach sketched above guarantees that, as in the MAP problem, the knowledge structures remain both isotropic and Quinean, i.e., everything still can be relevant to the outcome of the inference process and the candidate explanations remain sensitive to the entire belief system, as claimed in Section 1. For example, when new evidence arises (say, that we learn of a bank robbery where an innocent passerby was shot), the color of Mr. Jones's coat suddenly may become relevant to explaining his absence. We will elaborate on the *tractability* claim in Section 3.2 and on the *tracking truth* claim in Section 4.2.

Example 3. As in the previous example, we assume that in the ALARM network PCWP and BP have been observed to be high and the other observable variables take their non-alarming states. Furthermore, let us assume that we consider VTUB, SHNT, VLNG, VALV and LVV to be relevant intermediate variables, and VMCH, PVS, ACO2, CCHL, ERLO, STKV, HR, and ERCA to be irrelevant variables. The most *frugal* joint explanation for the diagnostic variables is still that HYP = true while all other diagnostic variables are negative: in 31% of the joint value assignments to these irrelevant intermediate variables, this is the most probable explanation. In 16% of the assignments 'all negative' is the most probable explanation, and in 24% of the assignments HYP = true and INT = one sided (one sided intubation, rather than normal) is the most probable explanation of the observations. If, in addition, we also consider VMCH, PVS, and STKV to be relevant, then every joint value assignment to ACO2, CCHL, ERLO, HR, and ERCA will have HYP = true as the most probable explanation for the observations. In other words, rather than marginalizing over these variables, we might have assigned just an arbitrary joint value assignment to these variables, decreasing the computational burden. If we had considered less intermediate variables to be relevant, this strategy may still often work, but has a chance of error, if we pick a sample for which a different explanation is the most probable one. We can decrease this error by taking more samples and take a majority vote.

Note that MFE is not *guaranteed* to give the MAP explanation, unless we marginalize over all intermediate variables (i.e., consider all variables to be relevant). When the set of irrelevant variables is non-empty, the most frugal explanation may differ from the MAP explanation, even when using a voting strategy based on *all* joint value assignments to the irrelevant intermediate variables, since both explanations are computed differently. Take for example the small network with two ternary variables H with values $\{h_1, h_2, h_3\}$ and I with values $\{i_1, i_2, i_3\}$, with I uniformly distributed and H conditioned on I as follows:

$$\begin{array}{lll}
\Pr(h_1 \mid i_1) = 0.4 & \Pr(h_2 \mid I = i_1) = 0.3 & \Pr(h_3 \mid i_1) = 0.3 \\
\Pr(h_1 \mid i_2) = 0.4 & \Pr(h_2 \mid I = i_2) = 0.3 & \Pr(h_3 \mid i_2) = 0.3 \\
\Pr(h_1 \mid i_3) = 0.1 & \Pr(h_2 \mid I = i_3) = 0.6 & \Pr(h_3 \mid i_3) = 0.3
\end{array}$$

Now, the most *probable* explanation of H, marginalized on I, would be $H = h_2$, but the most *frugal* explanation of H with irrelevant variable I would be $H = h_1$ as this is the most probable explanation for two out of three value assignments to I. Only in borderline cases MAP and MFE are guaranteed to give the same results independent of the number of samples taken and the partition in relevant and irrelevant intermediate variables. This will, for example, be the case when the MAP explanation has a probability of 1 and all the intermediate variables are uniformly distributed. In this case, every joint value assignment to any subset of the intermediate variables gives the MAP explanation as most frugal explanation.[4]

3.1. Relevance

Until now, we have quite liberally used the notion 'relevance'. It is important here to note that we consider the relevance of *intermediate* variables. This is in contrast with Shimony's well-known account [55] where relevance is a property of *explanation* variables, i.e., the relevance criterion partitions the non-observed variables in MAP variables—that are to be explained—and intermediate variables that do not need to be assigned a value in the explanation. In this paper we assume that the partition between the explanation variables $\mathbf{H}$ and the intermediate variables $\mathbf{I}$ is already made. However, in our model we again partition the intermediate variables $\mathbf{I}$ and perform full inference only on the *relevant* intermediate variables $\mathbf{I}^+$.

[4]We thank one of the anonymous reviewers for this observation.

It will be clear that the formal notion of (conditional) independence is too strong to capture relevance as we understand it: even if an intermediate variable is formally not independent of all the explanation variables, conditioned on the observed evidence variables, its influence may still be too small to have an impact on which explanation to select as the most probable as we saw in the previous sub-section. In contrast, we define relevance as a statistical property of an intermediate variable that is partly based on Druzdzel and Suermondt's [14] definition of relevance of variables in a Bayesian model, and partly on Wilson and Sperber's [65] relevance theory, and is related to the definition in [37]. According to Druzdzel and Suermondt a variable in a Bayesian model is relevant for a set $\mathbf{T}$ of variables, given an observation $\mathbf{E}$, if it is "needed to reason about the impact of observing $\mathbf{E}$ on $\mathbf{T}$" [14, p.60]. Our operationalization of "needed to reason" is inspired by Wilson and Sperber, who state that "an input is relevant to an individual when its processing in a context of available assumptions yields (...) a worthwhile difference to the individual's representation of the world" [65, p.608]. The term 'worthwhile difference' in this quote refers to the balance between the actual effects of processing that particular input and the effort required to do so. We therefore define the relevance of an intermediate variable as a *measure*, indicating how sensitive explanations are to changes in its value assignment. Informally, an intermediate variable I has a low relevance when there are only few possible worlds in which the most probable explanation changes when the value of I changes.[5]

Definition 4. Let $\mathcal{B} = (\mathbf{G}_{\mathcal{B}}, \Pr)$ be a Bayesian network partitioned into evidence nodes $\mathbf{E}$ with joint value assignment $\mathbf{e}$, intermediate nodes $\mathbf{I}$, and an explanation set $\mathbf{H}$. Let $I \in \mathbf{I}$, and let $\Omega(\mathbf{I} \setminus \{I\})$ denote the set of joint value assignments to the intermediate variables other than I. The *relevance* of I, denoted as $\mathcal{R}(I)$, is the fraction of joint value assignments $\mathbf{i}$ in $\Omega(\mathbf{I} \setminus \{I\})$ for which $\text{argmax}_{\mathbf{h}} \Pr(\mathbf{h}, \mathbf{e}, \mathbf{i}, i)$ is not identical for all $i \in \Omega(I)$.

As computing the relevance of a variable I is NP-hard, i.e., intractable in general (see the Appendix for a formal proof), we introduce the notion *estimated relevance of* I as a subjective assessment of $\mathcal{R}(I)$ which may or may not correspond to the actual value. Such a subjective assessment might be

[5]Note that the *size of the effect* on the probability distribution of $\mathbf{H}$ is not taken into account here, only that the distribution alters sufficiently enough for the most probable joint value assignment to 'flip over' to a different value.

based on heuristics, previous knowledge, or by approximating the relevance, e.g., by sampling a few instances of $\Omega(\mathbf{I}\backslash\{I\})$. Where confusion may arise, we will use the term *intrinsic relevance* to refer to the actual statistical property 'relevance' of I, in contrast to the subjective assessment thereof. Note that both intrinsic and estimated relevance of a variable are relative to a particular set of candidate explanations $\mathbf{H}$, and conditional on a particular observation, i.e., a value assignment $\mathbf{e}$ to the evidence nodes $\mathbf{E}$.

Example 5. Let, in the ALARM network, pulmonary capillary wedge pressure and blood pressure be high, and let all other observable variables take their non-alarming default values. The intrinsic relevance of the intermediate variables for the diagnosis is given in Figure 2.

When solving an MFE problem, we marginalize over the 'relevant intermediate variables'. This assumes some (subjective) threshold on the (estimated or intrinsic) relevance of the intermediate variables that determine which variables are considered to be relevant and which are considered to be irrelevant. For example, if the threshold would be 0.85 then only SHNT and LVV would be relevant intermediate variables in the ALARM network, but if the threshold would be 0.40 then also VMCH, VTUB, VLNG, VALV, and STKV would be relevant variables. That influences the results, as the distribution of MFE explanations tends to be flatter when less variables are marginalized over. With a threshold of 0.85 there are 24 explanations that are sometimes the most probable explanation, with the actual MAP explanation occurring most often (26%). With a threshold of 0.40 there are just three such explanations, with the MAP explanation occurring in 75% of the cases. Thus, the distribution of MFE explanations is typically more 'skewed' towards one explanation when more variables are considered to be relevant.

3.2. Complexity Analysis

To assess the computational complexity of MFE, we first define a decision variant.

Most Frugal Explanation (MFE)
Instance: A Bayesian network $\mathcal{B} = (\mathbf{G}_{\mathcal{B}}, \Pr)$, where $\mathbf{V}$ is partitioned into a set of evidence nodes $\mathbf{E}$ with a joint value assignment $\mathbf{e}$, an explanation set $\mathbf{H}$, a set of *relevant* intermediate variables $\mathbf{I}^{+}$, and a set of *irrelevant* intermediate variables $\mathbf{I}^{-}$; a rational number $0 \leq q < 1$ and an integer $0 \leq k < |\Omega(\mathbf{I}^{-})|$.

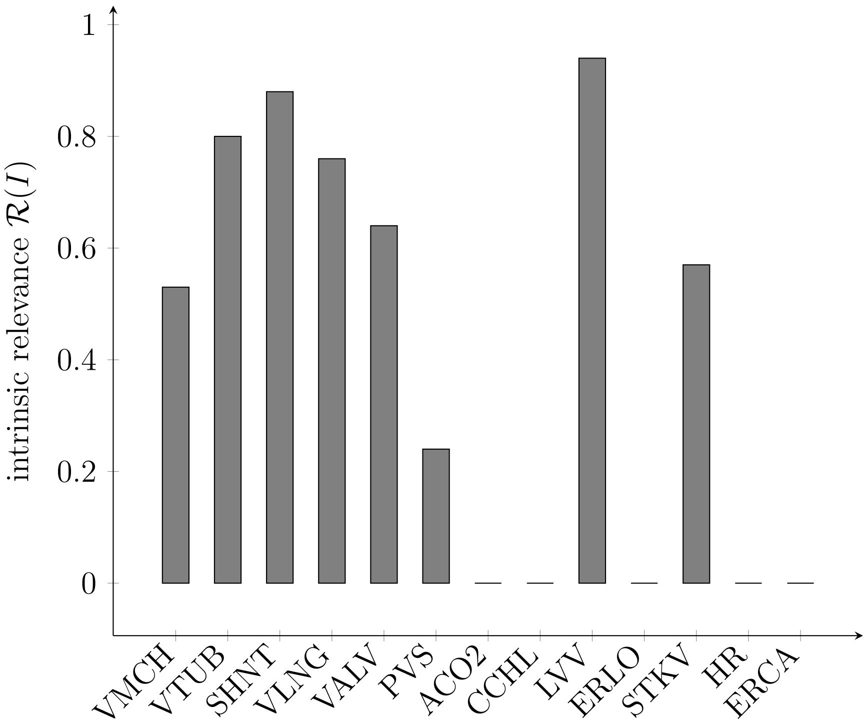

Figure 2: The intrinsic relevance of the intermediate variables of the ALARM network for the diagnostic variables given PCWP = TRUE and BP = TRUE. Note that the left ventricular end-diastolic blood volume (LVV) is highly relevant for the diagnosis, while the amount of catecholamines in the blood (CCHL) is irrelevant given these observations

---

Question: Is there a joint value assignment $\mathbf{h}$ to the nodes in $\mathbf{H}$ such that for more than k disjoint joint value assignments $\mathbf{i}$ to $\mathbf{I}^-$, $\Pr(\mathbf{h}, \mathbf{i}, \mathbf{e}) > q$?

It will be immediately clear that MFE is intractable, as it has the $\mathsf{NP}^{\mathsf{PP}}$-complete MAP [51] and MSE [35] problems as special cases for $\mathbf{I}^- = \varnothing$, respectively $\mathbf{I}^+ = \varnothing$. In this section we show that MFE happens to be even harder, viz., that it is $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$-complete, making it one of few real world-problems that are complete for that class[6]. The canonical SATISFIABILITY-

[6]Informally, one could imagine that for solving MFE one needs to counter *three* sources of complexity: selecting a joint value assignment out of potentially exponentially many

variant that is complete for this class is E-MajMajsat, defined as follows [61].

EMajMajsat
Instance: A Boolean formula ϕ whose n variables $x_1 \dots x_n$ are partitioned into three sets $\mathbf{E} = x_1 \dots x_k$, $\mathbf{M_1} = x_{k+1} \dots x_l$, and $\mathbf{M_2} = x_{l+1} \dots x_n$ for some numbers k, l with $1 \leq k \leq l \leq n$.
Question: Is there a truth assignment to the variables in $\mathbf{E}$ such that for the majority of truth assignments to the variables in $\mathbf{M_1}$ it holds, that the majority of truth assignments to the variables in $\mathbf{M_2}$ yield a satisfying truth instantiation to $\mathbf{E} \cup \mathbf{M_1} \cup \mathbf{M_2}$?

As an example, consider the formula $\phi_{\text{ex}} = x_1 \wedge (x_2 \vee x_3) \wedge (x_4 \vee x_5)$ with $\mathbf{E} = \{x_1\}$, $\mathbf{M_1} = \{x_2, x_3\}$ and $\mathbf{M_2} = \{x_4, x_5\}$. This is a *yes* example of E-MajMajsat: for $x_1 = $ TRUE, three out of four truth assignments to $\{x_2, x_3\}$ (all but $x_2 = x_3 = $ FALSE) are such that the majority of truth assignments to $\{x_4, x_5\}$ satisfy ϕ_{ex}.

To prove $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$-completeness of the MFE problem, we construct a Bayesian network $\mathcal{B}_\phi$ from an E-MajMajsat instance $(\phi, \mathbf{E}, \mathbf{M_1}, \mathbf{M_2})$. For each propositional variable x_i in ϕ, a binary stochastic variable X_i is added to $\mathcal{B}_\phi$, with uniformly distributed values TRUE and FALSE. These stochastic variables in $\mathcal{B}_\phi$ are three-partitioned into sets $\mathbf{X_E}$, $\mathbf{X_{M_1}}$, and $\mathbf{X_{M_2}}$ according to the partition of ϕ. For each logical operator in ϕ an additional binary variable in $\mathcal{B}_\phi$ is introduced, whose parents are the variables that correspond to the input of the operator, and whose conditional probability table is equal to the truth table of that operator. The variable associated with the top-level operator in ϕ is denoted as V_ϕ, the set of variables associated with the remaining operators is denoted as $\mathbf{Op}_\phi$. Figure 3 shows the graphical structure of the Bayesian network constructed for the example E-MajMajsat instance given above.

Theorem 6. MFE *is* $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$*-complete.*

candidate assignments to the explanation set; solving an inference problem over the variables in the set $\mathbf{I}^+$, and deciding upon a threshold of the joint value assignments to the set $\mathbf{I}^-$. While the 'selecting' aspect is typically associated with problems in NP, 'inference' and 'threshold testing' are typically associated with problems in PP. Hence, as these three sub-problems work on top of each other, the complexity class that corresponds to this problem is $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$.

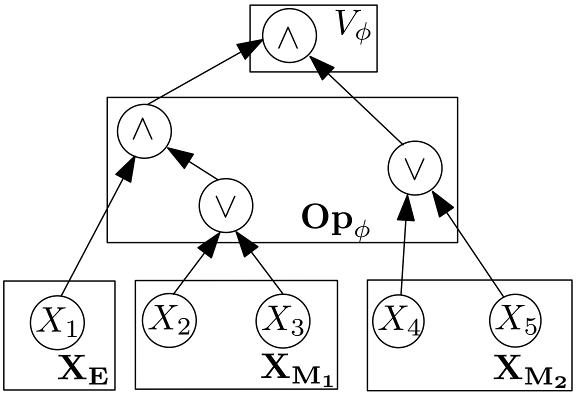

Figure 3: Example of the construction of $\mathcal{B}_{\phi_{\text{ex}}}$ for the Boolean formula $\phi_{\text{ex}} = x_1 \wedge (x_2 \vee x_3) \wedge (x_4 \vee x_5)$

Proof. Membership in $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$ follows from the following algorithm: non-deterministically guess a value assignment $\mathbf{h}$, and test whether there are at least k joint value assignments $\mathbf{i}^-$ to $\mathbf{I}^-$ such that $\Pr(\mathbf{h}, \mathbf{i}^-, \mathbf{e}) > q$. This inference problem can be decided (for given value assignment $\mathbf{h}$ and $\mathbf{i}^-$) using a PTM capable of deciding Inference (marginalizing over the variables in $\mathbf{I}^+$). We can decide whether there are at least k such joint value assignments $\mathbf{i}^-$ using an PTM capable of threshold counting. Thus, as both deciding Inference and threshold counting are PP-complete problems, we can solve this problem by augmenting an NTM with an oracle for $\mathsf{PP}^{\mathsf{PP}}$-complete problems; note that we cannot 'merge' both oracles as the 'threshold' oracle machine must accept inputs for which the Inference oracle answers 'no' as well as inputs for which the oracle answers 'yes'.

To prove $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$-hardness, we reduce MFE from E-MajMajsat. We fix $q = 1/2$ and $k = |\Omega(\mathbf{I}^-)|/2$. Let $(\phi, \mathbf{E}, \mathbf{M_1}, \mathbf{M_2})$ be an instance of E-MajMajsat and let $\mathcal{B}_\phi$ be the network constructed from that instance as shown above. We claim the following: If and only if there exists a satisfying solution to $(\phi, \mathbf{E}, \mathbf{M_1}, \mathbf{M_2})$, there is a joint value assignment to $\mathbf{x_E}$ such that $\Pr(V_\phi = \text{TRUE}, \mathbf{x_E}, \mathbf{x_{M_2}}) > 1/2$ for the majority of joint value assignments $\mathbf{x_{M_2}}$ to $\mathbf{X_{M_2}}$.

$\Rightarrow$ Let $(\phi, \mathbf{E}, \mathbf{M_1}, \mathbf{M_2})$ denote the satisfiable E-MajMajsat instance. Note that in $\mathcal{B}_\phi$ any particular joint value assignment $\mathbf{x_E} \cup \mathbf{x_{M_1}} \cup \mathbf{x_{M_2}}$ to

$\mathbf{X_E} \cup \mathbf{X_{M_1}} \cup \mathbf{X_{M_2}}$ yields $\Pr(V_\phi = \text{TRUE}, \mathbf{x_E}, \mathbf{x_{M_1}}, \mathbf{x_{M_2}}) = 1$, if and only if the corresponding truth assignment to $\mathbf{E} \cup \mathbf{M_1} \cup \mathbf{M_2}$ satisfies ϕ, and 0 otherwise. When marginalizing over $\mathbf{x_{M_1}}$ (and $\mathbf{Op}_\phi$) we thus have that a joint value assignment $\mathbf{x_E} \cup \mathbf{x_{M_2}}$ to $\mathbf{X_E} \cup \mathbf{X_{M_2}}$ yields $\Pr(V_\phi = \text{TRUE}, \mathbf{x_E}, \mathbf{x_{M_2}}) > 1/2$ if and only if the majority of truth assignments to $\mathbf{M_1}$, together with the given truth assignment to $\mathbf{E} \cup \mathbf{M_2}$, satisfy ϕ. Thus, given that this is the case for the majority of truth assignments to $\mathbf{M_2}$, we have that $\Pr(V_\phi = \text{TRUE}, \mathbf{x_E}, \mathbf{x_{M_2}}) > 1/2$ for the majority of joint value assignments $\mathbf{x_{M_2}}$ to $\mathbf{X_{M_2}}$. We conclude that the corresponding instance $(\mathcal{B}_\phi, V_\phi = \text{TRUE}, \mathbf{X_E}, \mathbf{X_{M_1}} \cup \mathbf{Op}_\phi, \mathbf{X_{M_2}}, 1/2, |\Omega(\mathbf{X_{M_2}})|/2)$ of MFE is satisfiable.

$\Leftarrow$ Let $(\mathcal{B}_\phi, V_\phi = \text{TRUE}, \mathbf{X_E}, \mathbf{X_{M_1}} \cup \mathbf{Op}_\phi, \mathbf{X_{M_2}}, 1/2, |\Omega(\mathbf{X_{M_2}})|/2)$ be a satisfiable instance of MFE, i.e., there exists a joint value assignment $\mathbf{x_E}$ to $\mathbf{X_E}$ such that for the majority of joint value assignments $\mathbf{x_{M_2}}$ to $\mathbf{X_{M_2}}$, $\Pr(V_\phi = \text{TRUE}, \mathbf{x_E}, \mathbf{x_{M_2}}) > 1/2$. For each of these assignments $\mathbf{x_{M_2}}$ to $\mathbf{X_{M_2}}$, $\Pr(V_\phi = \text{TRUE}, \mathbf{x_E}, \mathbf{x_{M_2}}) > 1/2$ if and only if the majority of joint value assignments $\mathbf{x_{M_1}}$ to $\mathbf{X_{M_1}}$ satisfy ϕ.

Since the reduction can be done in polynomial time, this proves that MFE is $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$-complete. □

Given the intractability of MFE for unconstrained domains, it may not be clear how MFE as a heuristic mechanism for Bayesian abduction can scale up to task situations of real-world complexity. One approach may be to seek to approximate MFE, rather than to compute it exactly. Unfortunately, *approximating* MFE is NP-hard as well. Given that MFE has both MAP and MSE as special cases (for $\mathbf{I}^- = \varnothing$, respectively $\mathbf{I}^+ = \varnothing$), it is intractable to infer an explanation that has a probability that is close to optimal [51], that is similar to the most frugal explanation [40], or that is likely to be the most frugal explanation with a bounded margin of error [42]. By and of itself, for unconstrained domains, approximation of MFE does not buy tractability [43].

3.3. Parameterized Complexity

An alternative approach to ensure computational tractability is to study how the complexity of MFE depends on situational constraints. This approach has firm roots in the theory of parameterized complexity as described in Section 2. Building on known fixed parameter tractability results for MAP

[36] and MSE [42], we will consider the parameters *treewidth* and *cardinality* of the Bayesian network, the *size* of $\mathbf{I}^+$, and a *decisiveness* measure on the probability distribution. An overview is given in Table 1.

Parameter	Description
Treewidth (t)	A measure on the network topology [see, e.g., 3].
Cardinality (c)	The maximum number of values any variable can take.
#Relevants ($\lvert\mathbf{I}^+\rvert$)	The number of relevant intermediate variables that we marginalize over.
Decisiveness (d)	A measure on the probability distribution [42], denoting the probability that for a given evidence set $\mathbf{E}$ with evidence $\mathbf{e}$ and explanation set $\mathbf{H}$, two random joint value assignments $\mathbf{i_1}$ and $\mathbf{i_2}$ to the irrelevant variables $\mathbf{I}^-$ would yield the same most probable explanations. Decisiveness is high if a robust majority of the joint value assignments to $\mathbf{I}^-$ yields a particular most probable explanation.

Table 1: Overview of parameters for MFE.

For $\mathbf{I}^+ = \varnothing$, MAP can be solved in $O(c^t \cdot n)$ for a network with n variables, and since $\Pr(X = x) = \sum_{y \in \Omega(Y)} \Pr(X = x, Y = y)$, we have that MAP can be solved in $O(c^t \cdot c^{|\mathbf{I}^+|} \cdot n)$. Note that even when we can tractably decide upon the most probable explanation for a given joint value assignment $\mathbf{i}$ to $\mathbf{I}^-$ (i.e., when c, t, and $|\mathbf{I}^+|$ are bounded) we still need to test at least $\lfloor c^{|\mathbf{I}^-|}/2 \rfloor + 1$ joint value assignments to $|\mathbf{I}^-|$ to decide MFE exactly, even when $d = 1$. However, in that case we can tractably find an explanation that is *very likely* to be the MFE if d is close to 1. Consider the following algorithm for MFE (adapted from [35]):

Algorithm 1 Compute the Most Frugal Explanation

Sampled-MFE($\mathcal{B}, \mathbf{H}, \mathbf{I}^+, \mathbf{I}^-, \mathbf{e}, N$)

for $n = 1$ to N **do**
  Choose $\mathbf{i} \in \mathbf{I}^-$ at random
  Determine $\mathbf{h} = \text{argmax}_{\mathbf{h}} \Pr(\mathbf{H} = \mathbf{h}, \mathbf{i}, \mathbf{e})$
  Collate the joint value assignments $\mathbf{h}$
end for
Decide upon the joint value assignment $\mathbf{h}_{\text{maj}}$ that was picked most often
return $\mathbf{h}_{\text{maj}}$

This randomized algorithm repeatedly picks a joint value assignment $\mathbf{i} \in \mathbf{I}^-$ at random, determines the most probable explanation, and at the end decides upon which explanation was found most often. Due to its stochastic nature, this algorithm is not guaranteed to give correct answers all the time. However, the error margin ϵ can be made sufficiently low by choosing N large enough. If there are only two competing most probable explanations, the threshold value of N can be computed using the *Chernoff bound*: $N \geq \frac{1}{(p - 1/2)^2} \ln 1/\sqrt{\epsilon}$ (more sophisticated methods are to be used to compute or approximate N when there are more than two competing explanations). Assume we require an error margin of less than 0.1, then the number of repeats depends on the probability p of picking a joint value assignment $\mathbf{i}$ for which $\mathbf{h}_{\text{maj}}$ is the most probable explanation. This probability corresponds to the *decisiveness* parameter d that was introduced in Table 1. If decisiveness is high (say $d = 0.85$), then N can be fairly low ($N \geq 10$), however, if the distribution of explanations is very flat, and consequently, decisiveness is low, then an exponential number of repetitions is needed.

If d is bounded (i.e., larger than a particular fixed threshold) we thus need only polynomially many repetitions to obtain any constant error rate. When in addition determining the most probable explanation is easy—in particular, when the treewidth and cardinality of $\mathcal{B}$ are low and there are few relevant variables in the set $\mathbf{I}^+$—the algorithm thus runs in polynomial time, and thus MFE can be decided in polynomial time, with a small possibility of error.

3.4. Discussion

In the previous subsections we showed that MFE is intractable in general, both to compute exactly and to approximate, yet can be tractably approximated (with a so-called expectation-approximation [42]) when the treewidth of the network is low, the cardinality of the variables is small, the number of relevant intermediate variables is low, *and* the probability distribution for a given explanation set $\mathbf{H}$, evidence $\mathbf{e}$ and relevant intermediate variables $\mathbf{I}^+$ is fairly decisive, i.e., skewed towards a single MFE explanation. We also know that MAP can be tractably computed exactly[7] when the treewidth of the network is low, the cardinality of the variables is small, and either the MAP explanation has a high probability, or the total number of intermediate variables is low [36]. How do these constraints compare to each other?

For MAP, the constraint on the total number of intermediate variables seems implausible. In real-world knowledge structures there are many intermediate variables, and while only some of them may contribute to the MAP explanation, we still need to marginalize over all of them to compute MAP. Likewise, when there are many candidate hypotheses, it is not obvious that the most probable one has a high (i.e., close to 1) probability. Note that the actual fixed-parameter tractable algorithm [4, 36] has a running time with $\frac{\log p}{\log 1-p}$ in the exponent, where p denotes the probability of the MAP explanation. This exponent quickly grows with decreasing p. Furthermore, treewidth and cardinality actually refer to the treewidth of the *reduced* junction tree, where observed variables are absorbed in the cliques. Given that we sample over the set $\mathbf{I}^-$ in MFE, but not in MAP, both parameters (treewidth and cardinality) will typically have much lower values in MFE as compared to MAP. That is, it is more plausible that these constraints are met in MFE than that they are met in MAP.

Given the theoretical considerations in [14] it seems plausible that the *decisiveness* constraint is met in many practical situations. Surely, one could argue that the fixed parameter tractability of MFE is misguided, as the set of candidate hypotheses and the observations are given in the input of the formal problem, and it is known beforehand what the relevant variables are. Thus, the problem of finding candidate hypotheses, the problem of deciding what counts as evidence, and the problem of deciding which variables are relevant are left out of the problem definition. We acknowledge that this is indeed the case, and that the problem of non-demonstrative inference is much broader than 'merely' inferring the best explanation out of a set of

[7] There are to the best of our knowledge no stronger (or even *different*) fixed parameter tractable results for *approximate* MAP than the results listed above for exact computations.

candidate explanations [39]; yet, this is no different for MAP, at least when it comes to deciding upon the candidate hypotheses and the observations. With respect to the partition between irrelevant and relevant intermediate variables we will show in Section 4 that MFE is fairly robust: including even a few variables with a high intrinsic relevance will suffice to find relatively good MFE explanations.

4. Simulations

In Section 3 we illustrated, using the ALARM example, that computing MFE can give similar results as when MAP is computed, while requiring less variables to be marginalized over. In this section, we will simulate MFE on random graphs to obtain empirical results to support that claim. We will also illustrate that, in order to obtain a good explanation using only few samples, the decisiveness of the probability distribution indeed must be high. Finally we show how MFE behaves under various scenarios where the intrinsic and estimated relevance of the intermediate variables (i.e., the actual relevance and the subjective assessment thereof) do not match. As the goal of these simulations is to explore how MFE behaves under scenarios that can be considered either natural (occurring in real-world networks) or artificial, we use randomly generated networks, rather than an existing set of benchmark networks, like the ALARM network, in our simulations.

4.1. Method

We generated 100 random Bayesian networks, each consisting of 40 variables, using the (second) method described in [51]. Each variable had either two, three, or four possible values, and the in-degree of the nodes was limited to five. With each variable, a random conditional probability distribution was associated. We randomly selected five explanation variables and five evidence variables, and set a random joint value assignment to the evidence variables. Given the variation on the cardinality of the variables, the number of candidate joint value assignments to the explanation variables could vary from 2^5 to 4^5; in practice, it ranged from 48 to 576 (mean 208.5, standard deviation 107.4). See also the on-line supplementary materials: `http://www.dcc.ru.nl/~johank/MFE/`.

Using the Bayes Net Toolbox for MATLAB [46] we computed, for each network, the posterior distribution over the explanation variables, approximated the relevance of each intermediate variable, and approximated the

MFE distribution under various conditions. The results presented below are based on 91 random networks. The MATLAB software was unable to compute the MAP of seven networks due to memory limitations, and the results of two networks were lost due to hardware failure. In Figures 4 and 5 some typical results are given for illustrative purposes.

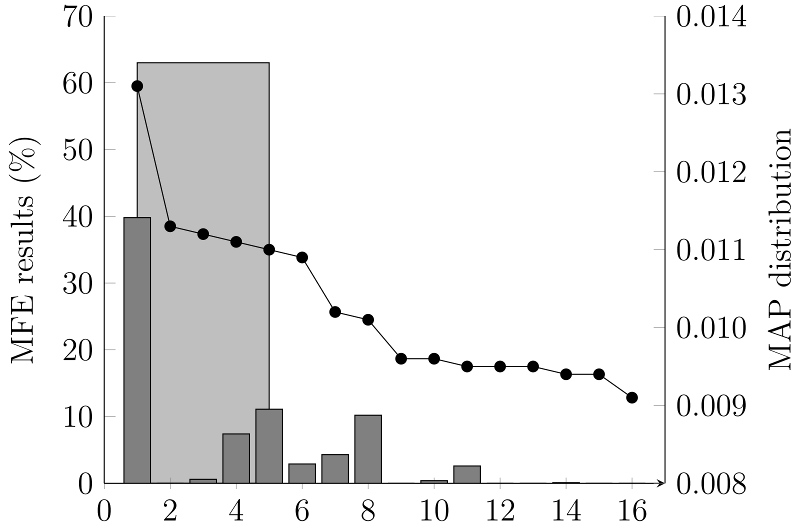

Figure 4: MAP distribution and MFE results for the 16 most probable joint value assignments of one of the random networks (#99) for a particular set of relevant intermediate variables, using 1000 samples. The light gray bar denotes the cumulative MFE result of the five most probable joint value assignments. Note that the most probable joint value assignment (which has a probability of 0.0131) is also the most frugal explanation, as it is the MAP for about 40% of the joint value assignments to the irrelevant intermediate variables. The 'second-best MAP', while it has a relative high posterior probability, is *always 'second-best'*: there are no joint value assignments to the irrelevant intermediate variables in which this particular explanation has the highest probability. There *are* other explanations, with a lower posterior probability, that become the most probable explanation for some particular value assignments to these irrelevant intermediate variables. Note that in this situation there is no error as the most probable and most frugal explanation are identical.

4.2. Tracking Truth

We compared the MAP explanation with the MFE explanation using 100 samples of the irrelevant variables, varying the I^+/I^- partition. In particular

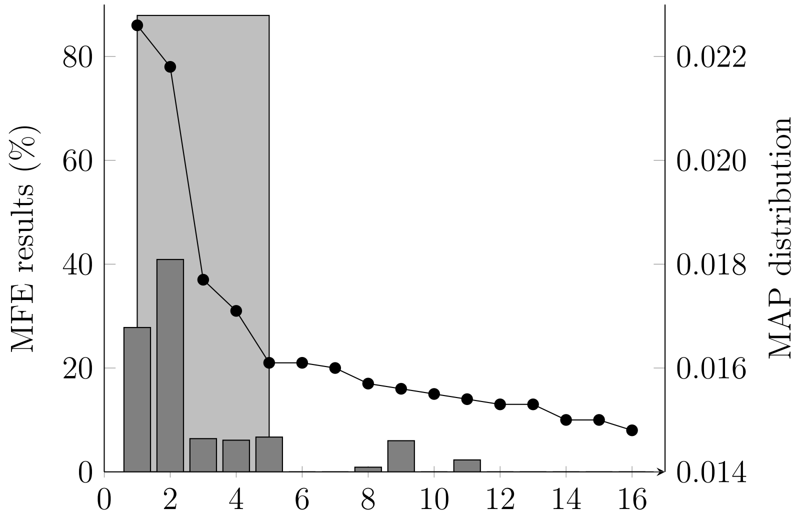

Figure 5: A similar plot as in Figure 4, but in this random network (#68) the most frugal explanation is the second most probable explanation, yielding a difference between the 'marginalizing' and the 'sampling' approach. Note, however, that both explanations are almost as good: they differ in a single variable, and the probability ratio is 0.965, meaning that the probability of the most frugal explanation is only slightly lower than the probability of the most probable explanation.

---

we compared the explanations where all variables are deemed irrelevant ($I^+ = \varnothing$), where I^+ consisted of the five intermediate variables with the highest relevance, and where I^+ consisted of the intermediate variables that have a relevance of more than 0.00, 0.05, 0.10, 0.25, respectively 0.50. To assess how similar the most frugal explanations are to the MAP results, we used three different error measures: (1) the structural deviation from MAP (how many variables have different values, i.e., the Hamming distance between the MFE and MAP explanations), (2) the rank k of the MFE explanation, indicating that the MFE explanation is the k-th MAP instead of the most probable explanation, and (3) the ratio of the MFE probability and the MAP probability, indicating the proportion of probability mass that was allocated to the MFE explanation.

Furthermore, we estimated how often the MFE was picked relative to other explanations, indicating how likely it is that a singleton sample over the irrelevant variables would yield this particular explanation. This yields a measure on how many samples are needed to arrive at a confident decision. Lastly, we estimated the likelihood of picking the MAP explanation and one of the five most probable explanations using a single sample. This indicates how likely it is that an arbitrary singleton sample will yield an explanation with the maximum, respectively a relatively high, posterior probability.

The results are summarized in Table 2 and Figure 6. The scatter plots in Figure 6 illustrate the spread of the errors along different networks. In general one can conclude that MFE explanations are reasonably close to the MAP explanations, when there is marginalization over those variables that are 'sufficiently relevant'. From the results it follows that including the five most relevant variables gives fairly good results, and that including variables that have a relevance of less than 0.25 does not significantly improve the average MFE results. Including no relevant variables at all (i.e., computing the Most Simple Explanation [35]) gives considerably worse results, however.

Cond.	I^+ size	ratio	rank	dist.	% MFE	% MAP	% 5-MAP
None	0	0.66	25.90	2.05	0.08	0.03	0.14
Best 5	5	0.82	10.73	1.30	0.13	0.08	0.27
> 0.50	11.32	0.87	5.36	0.87	0.25	0.17	0.46
> 0.25	14.93	0.91	4.59	0.79	0.38	0.25	0.58
> 0.10	15.79	0.91	5.56	0.81	0.39	0.25	0.60
> 0.05	15.99	0.91	6.09	0.75	0.41	0.27	0.60
> 0.00	16.35	0.92	4.12	0.75	0.41	0.26	0.61

Table 2: Overview of simulation results. In this simulation the partition between relevant and irrelevant variables was varied and ranged from 'none' (all variables are irrelevant), 'best 5' (the five variables with the highest relevance are deemed relevant, to a relevance threshold between 0.50 and 0.00, yielding an average I^+ size between 11.32 and 16.35.

4.3. Number of Samples

As shown in Section 3.3, approximating the MFE (i.e., finding the explanation which is very likely the MFE) can be done by sampling, where the number of samples needed to guarantee a particular confidence level is related to the decisiveness of the network. When decisiveness is low, and consequently the MFE distribution is flat (many competing explanations, none of which has a high probability of being the most probable explanation for a random joint value assignment to the irrelevant intermediate variables), we

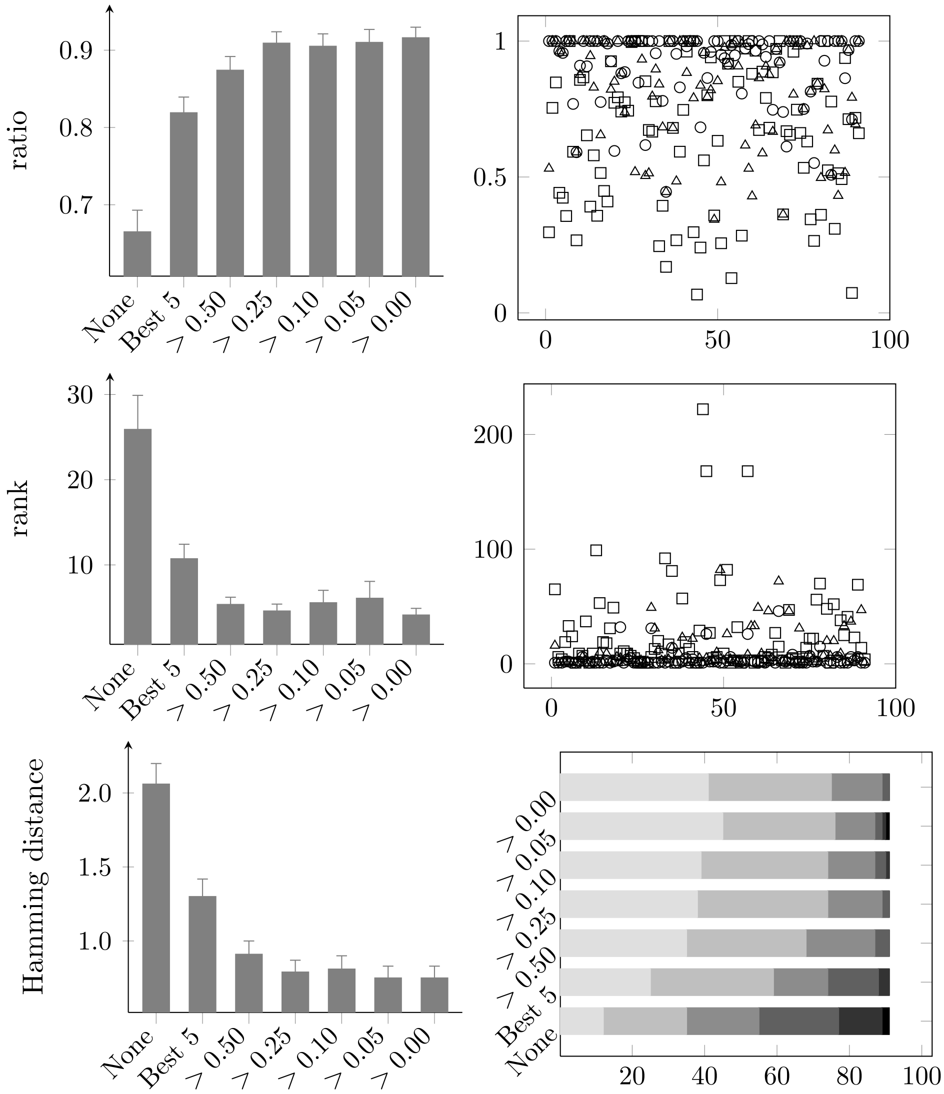

Figure 6: On the left: Three error indicators of MFE versus MAP explanations: The ratio between their probabilities, rank of the MFE explanation, and Hamming distance between MFE and MAP for various I^+/I^- settings. On the right: Scatter plots of ratio and rank, and stacked box plot for Hamming distance. In the scatter plots, results of all random networks are shown, for the conditions where all variables are irrelevant ('None', square), the five variables with the highest relevancy were deemded relevant ('best 5', triangle) and where all variables with non-zero relevancy where relevant ('$>$ 0.00', circle). The stacked box plot illustrates the distribution of the Hamming distance between MFE and MAP explanation, where darker colors indicate a higher Hamming distance. Error bars indicate standard error of the mean.

need much more samples to make confident decisions. This is illustrated by the following figures. In Figure 7 we see a typical result for a random network which is highly skewed towards a singleton explanation, and in Figure 8 the results of a random network with a low decisiveness are shown.

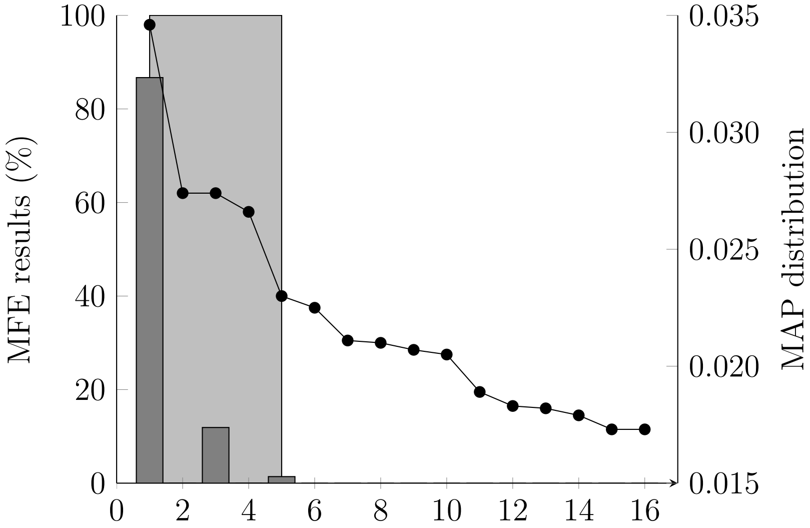

Figure 7: This plot shows part of the MAP distribution and MFE results using 1000 samples for a random network (#93) with a very steep distribution of the MFE explanations. This network is strongly skewed towards the most probable explanation which is picked in 83% of the samples, so that an arbitrary singleton sample is quite likely to be the MFE; we can be guaranteed to obtain the most frugal explanation with 95% confidence by generating thirteen samples and decide which explanation is most often picked. Even a single sample is guaranteed to correspond to one of the five most probable examples.

However, even when there is no explanation which stands out, the sampling algorithm can still give good results. In Figure 9 we show a typical result when there are *a few* competing explanations that all have a relatively high probability. While it may take many samples to decide on which of them is the MFE, we still can be quite sure that a singleton sample of the irrelevant intermediate variables would yield *one of them* as the most probable explanation; here, sampling seems like a reasonable strategy to obtain an explanation that is likely to have a reasonably high probability.

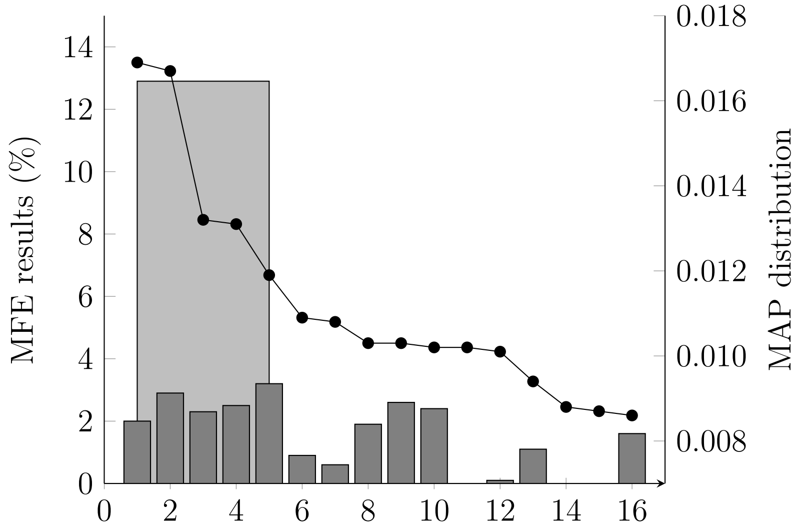

Figure 8: This plot shows part of the MAP distribution and MFE results using 1000 samples for a random network (#89) with a very flat distribution of the MFE explanations. No explanation really stands out; the most frugal explanation being picked in just over 3% of the samples. In this network, that is not at all skewed towards any particular explanation, an arbitrary sample can have a low posterior probability, and we will need a massive number of samples to decide with reasonable confidence about which explanation is the MFE.

4.4. Other parameters

Obviously, the $\mathbf{I}^+/\mathbf{I}^-$ partition influences the quality of the MFE solution in terms of the three error measures introduced in Section 4.2. We also investigated whether the size of the hypothesis space, the number of relevant variables, or the probability of the most probable explanation influences this quality. First we observe that these parameters are not independent. There is a strong negative correlation (-.65) between the size of the explanation set and the probability of the most probable explanation. This can be explained by the random nature of the networks and the probability distribution they capture: on average, if there are more candidate explanations in the explanation set, the average probability of each of them is lower, and so it is expected that the average probability of the most probable explanation is also lower. The results of the correlation analysis are shown in Table 3, and can be summarized as follows. Neither explanation set size, intrinsic relevance, or

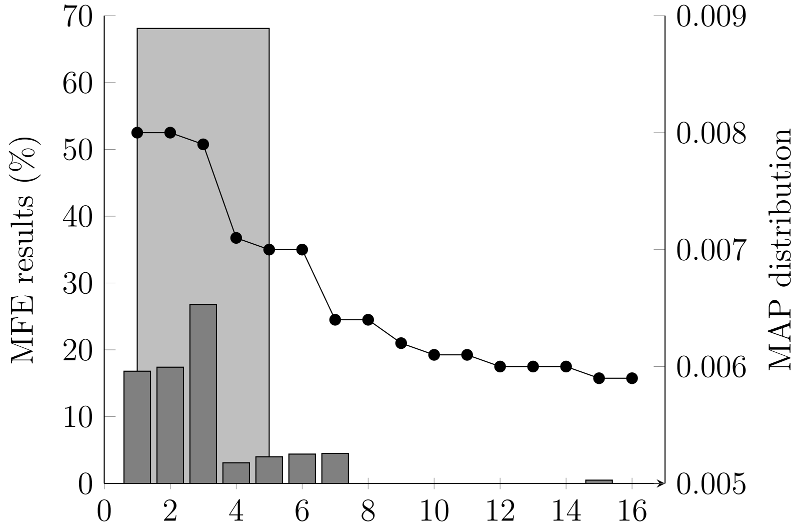

Figure 9: This plot shows part of the MAP distribution and MFE results using 1000 samples for a random network (#70) where three explanations are often picked as the most probable, and quite some samples are needed to decide on the most frugal explanation with confidence. However, since one of these three (almost equally probable) most probable explanations is picked in 61% of the samples, we can expect that few samples, possibly just a singleton sample, may return a quite good explanation.

probability of the most probable explanation (MPE) correlates with the ratio between probability of MPE and probability of MFE. There is a weak correlation between explanation set size and rank, and a weak negative correlation between probability of MPE and rank: the bigger the explanation size, the larger the average rank k. Neither explanation set size, intrinsic relevance, or probability of MPE correlates (or correlates only very weakly) with distance errors.

4.5. Wrong judgments

Obviously, taking more intermediate variables into account (i.e., considering more variables to be relevant) helps to obtain better results; still, we can make reasonable good inferences using only the five most relevant intermediate variables. But what if ones subjective assessment of what is relevant does not match the intrinsic relevance of these variables? Figure 10 illustrates what typically happens when there is a mismatch between intrinsic

Cond.	explanation set size			intrinsic relevance			probability of MPE		
	ratio	rank	dist.	ratio	rank	dist.	ratio	rank	dist.
MSE	$-.01$	.15	.15	$-.09$	.02	.18	$-.11$	$-.23^{*}$	$-.20$
Best 5	$-.16$	$.22^{*}$	$.27^{*}$	.13	.18	.07	$-.15$	$-.35^{**}$	$-.40^{**}$
> 0.50	.08	.12	.02	$-.11$	.01	.18	$-.04$	$-.17$	$-.16$
> 0.25	$-.09$	$.24^{*}$	.12	$-.11$	.05	$-.02$	.06	$-.22^{*}$	$-.12$
> 0.10	$-.10$	$.26^{*}$	$.21^{*}$	$-.08$	.01	$-.06$	.05	$-.17$	$-.18$
> 0.05	$-.08$	$.22^{*}$	.10	$-.08$	.01	$-.02$	.02	$-.13$	$-.11$
> 0.00	.06	.17	.03	$-.20$	.11	.01	$-.01$	$-.19$	$-.03$

Table 3: Overview of correlations (Pearson's r) with significance levels. * indicates significance at the $p < .05$ level, ** indicates significance at the $p < .01$ level

and estimated relevance. Here we plotted the results of the > 0.00 (top left) and Best 5 (bottom right) conditions, as well as some conditions in which there is a mismatch between intrinsic and expected relevance. In particular, we omitted the two (top right), five (middle left), ten (middle right), respectively fifteen (bottom left) most relevant variables.

This example illustrates a graceful degradation of the results, especially when the cumulative results of the five most probable joint value assignments are compared. Observe that including the twenty-five *least* relevant variables leads to comparable results as including the five *most* relevant variables. Clearly, it helps to know what is relevant, yet there is margin for error.

4.6. Discussion

The simulation results, as illustrated by Table 2 and Figure 6, clearly show that MFE 'tracks truth' quite well, even when only part of the relevant intermediate variables are taken into account. However, when more intermediate variables are marginalized over, we can be more confident of the results. In these cases the distribution of explanations typically is narrower and it is more likely that a majority vote using few samples, or even a singleton sample, results in an explanation that is close to the most probable explanation. The simulations also indicate that indeed the probability distribution must be skewed towards one or a few explanations for obtaining acceptable results with few samples - and that indeed many distributions *are* skewed, even in completely random networks.

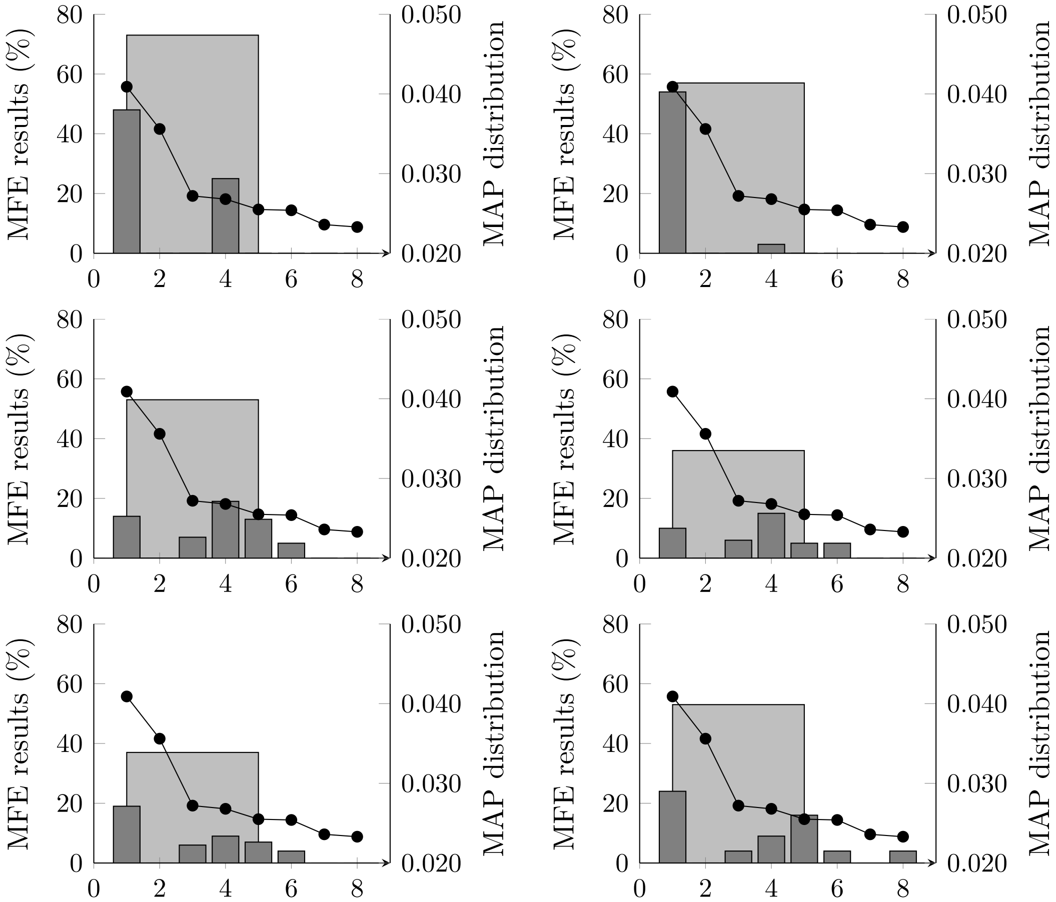

Figure 10: This plot shows part of the MAP distribution and MFE results of a random network (#78) with different partitions of the intermediate variables, where the subjective assessment that yields the partition may not match the actual relevance of the variables. Shown are the results when all variables with non-zero relevancy are deemed relevant (top left, 19 variables in $\mathbf{I}^+$), all *but* the two most relevant variables (top right, 28 variables in $\mathbf{I}^+$), all *but* the five most relevant variables (middle left, 25 variables in $\mathbf{I}^+$), all *but* the ten most relevant variables (middle right, 20 variables in $\mathbf{I}^+$), only the fifteen *least* relevant variables (bottom left, 15 variables in $\mathbf{I}^+$), and only the five *most* relevant variables (bottom right, 5 variables in $\mathbf{I}^+$).

5. Conclusion

In this paper we proposed Most Frugal Explanation (MFE) as a tractable heuristic alternative to (approximate) MAP for deciding upon the best explanation in Bayesian networks. While the MFE problem is intractable in general—its decision variant is $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$-complete, and thus even harder than the $\mathsf{NP}^{\mathsf{PP}}$-complete MAP problem [51], the $\mathsf{PP}^{\mathsf{PP}}$-complete Same-Decision Probability problem [9], or the $\mathsf{P}^{\mathsf{PP}^{\mathsf{PP}}}$-complete k-th MAP problem [41]—it can be tractably approximated under situational constraints that are arguably more realistic in large real-world applications than the constraints that are needed to render MAP (fixed-parameter) tractable. Notably, the $\{c, tw, 1-p\}$-fixed-parameter tractable algorithm for MAP [4] has a running time with $\frac{\log p}{\log 1-p}$ in the exponent. In the random networks used in the simulations, the *average* probability of the most probable explanation was 0.0245, which would yield an unpractical exponent of $\frac{\log 0.0245}{\log 0.9755} \approx 150$. In contrast, even when only half of the total set of intermediate variables are considered as relevant, for an arbitrary sample over the rest of the intermediate variables we will find the MFE in about 40% of the cases, and an explanation that is one of the five best in about 60% of the cases.

In future work we wish to investigate the possible explanatory power of MFE in cognitive science. In recent years it has been proposed that human cognizers make decisions using (Bayesian) sampling [31, 57, 62] and approximate Bayesian inferences using exemplars [54]; studies show that we have a hard time solving problems with many relevant aspects [20]. The parameterized complexity results of the MFE framework may theoretically explain why such approaches work fine in practice and under what conditions the limits of our cognitive capacities are reached.

6. Acknowledgments

The author wishes to thank the members of the *Theoretical Cognitive Science* group at the Donders Center for Cognition for useful discussions and comments on earlier versions of this paper, and the anonymous reviewers that gave valuable suggestions for improvement. He is in particular indebted to Todd Wareham for suggesting the term “Most Frugal Explanations” to denote the problem of finding an explanation for observations without taking care of everything that is only marginally relevant. A previous shorter version of this paper appeared in the Benelux Conference on AI [38].

7. Vitae

Johan Kwisthout is a postdoctoral researcher at the Donders Institute for Brain, Cognition and Behaviour, and lecturer in the Department of Artificial Intelligence, at the Raboud University Nijmegen. His research interests are reasoning under uncertainty, computational complexity, and the application of both fields of study in (theoretical) cognitive (neuro-)science. In particular, he is interested in topics related to (Bayesian) abduction and relevance, both from a philosophical, cognitive, and computational perspective.

Appendix: Computing relevance is NP-hard

In Definition 4 we formally defined the intrinsic relevance of an intermediate variable as a measure indicating the sensitivity of explanations to its value. We here show that computing the intrinsic relevance of such a variable is NP-hard. The decision problem used in this proof is defined as follows:

Intrinsic Relevance
Instance: A Bayesian network $\mathcal{B} = (\mathbf{G}_{\mathcal{B}}, \Pr)$, where $\mathbf{V}$ is partitioned into evidence variables $\mathbf{E}$ with joint value assignment $\mathbf{e}$, explanation variables $\mathbf{H}$, and intermediate variables $\mathbf{I}$, and a designated variable $I \in \mathbf{I}$.
Question: Is the *intrinsic relevance* $\mathcal{R}(I) > 0$?

We reduce from the following NP-complete decision problem [37]:

IsA-Relevant Variable
Instance: A Boolean formula ϕ with n variables, describing the characteristic function $\mathbf{1}_\phi : \{\text{FALSE}, \text{TRUE}\}^n \to \{1, 0\}$, designated variable $x_r \in \phi$.
Question: Is x_r a relevant variable in ϕ, that is, is $\mathbf{1}_\phi(x_r = \text{TRUE}) \neq \mathbf{1}_\phi(x_r = \text{FALSE})$?

Here, the characteristic function $\mathbf{1}_\phi$ of a Boolean formula ϕ maps truth assignments to ϕ to $\{0, 1\}$, such that $\mathbf{1}_\phi(x) = 1$ if and only if x denotes a satisfying truth assignment to ϕ, and 0 otherwise. We will use the formula $\phi_{\text{ex}} = \neg(x_1 \vee x_2) \wedge x_3$ as a running example, where x_3 is the variable of interest. Note that x_3 is relevant, since for $x_1 = x_2 = \text{FALSE}$, $\mathbf{1}_\phi(x_3 = \text{TRUE}) \neq \mathbf{1}_\phi(x_3 = \text{FALSE})$.

We construct a Bayesian network $\mathcal{B}_\phi$ from ϕ as follows. For each propositional variable $x_i \in \phi$ we add a binary stochastic variable $X_i \in \mathcal{B}_\phi$ with uniformly distributed values TRUE and FALSE. We add an additional binary variable X_r^T with observed value TRUE. For each logical operator o_j in ϕ, we add *two* binary stochastic variables O_j and O_j^T in $\mathcal{B}_\phi$. The parents of the variables O_j are the variables O_k that represent the sub-formulas bound by O_j; in case such a sub-formula is a literal, the corresponding parent is a variable X_i. In contrast, the parents of the variables O_j^T are the variables O_k^T (for sub-formula), X_i (for literals *except* x_r), respectively X_r^T (for the literal x_r). The variables corresponding with the top-level operator in ϕ are denoted with V_ϕ, respectively V_ϕ^T.

Furthermore, an additional binary variable C is introduced in $\mathcal{B}_\phi$, acting as 'comparator' variable. C has both V_ϕ and V_ϕ^T as parents and conditional probability $\Pr(C = \text{TRUE} \mid V_\phi, V_\phi^T) = 1$ if $V_\phi \neq V_\phi^T$ and $\Pr(C = \text{TRUE} \mid V_\phi, V_\phi^T) = 0$ if $V_\phi = V_\phi^T$. An example of this construction is given in Figure 11 for the formula ϕ_{ex}. We set $\mathbf{H} = C$, $\mathbf{E} = X_r^T$, and $I = X_r$.

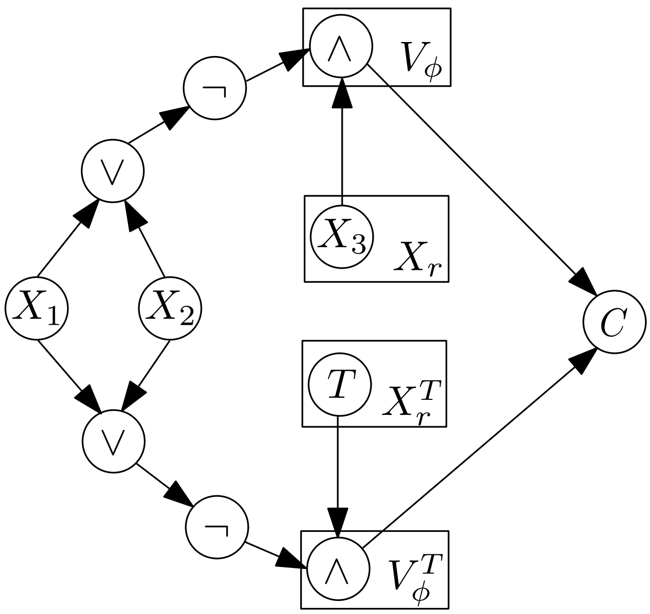

Figure 11: Example of the construction of $\mathcal{B}_{\phi_{\text{ex}}}$ for the formula $\phi_{\text{ex}} = \neg(x_1 \vee x_2) \wedge x_3$

Theorem 7. INTRINSIC RELEVANCE *is* NP*-complete.*

Proof. Membership in NP follows from the following polynomial-time verifying algorithm for *yes*-instances: given a suitable joint value assignment $\mathbf{i}$ to $\mathbf{I} \setminus \{I\}$ and assignments i_1, i_2 to I, we can easily check that $\text{argmax}_{\mathbf{h}} \Pr(\mathbf{h}, \mathbf{e}, \mathbf{i}, I = i_1) \neq \text{argmax}_{\mathbf{h}} \Pr(\mathbf{h}, \mathbf{e}, \mathbf{i}, I = i_2)$, and thus that $\mathcal{R}(I) > 0$.

To prove NP-hardness, we reduce ISA-RELEVANT VARIABLE to INTRINSIC RELEVANCE. Let (ϕ, x_r) be an instance of ISA-RELEVANT VARIABLE. From (ϕ, x_r), we construct $(\mathcal{B}_\phi, I)$ as described above. If (ϕ, x_r) is a *yes*-instance of ISA-RELEVANT VARIABLE, then the characteristic function $\mathbf{1}_\phi$ is not identical for $x_r = \text{FALSE}$ and $x_r = \text{TRUE}$. In other words, there is at least one truth assignment $\mathbf{t}$ to the variables in $\phi \setminus \{x_r\}$ such that either $\mathbf{t} \cup \{x_r = \text{TRUE}\}$ is satisfying ϕ and $\mathbf{t} \cup \{x_r = \text{FALSE}\}$ is not satisfying ϕ, or vice versa. Let $\mathbf{i}$ be the joint value assignment to $\mathbf{I} \setminus \{X_r\}$ that corresponds to the truth assignment $\mathbf{t}$, and in addition fixes the values of the operator variables O_j^T and O_j according to their (deterministic) conditional probability tables. Now, we have that for the truth assignment $X_r = \text{TRUE}$, $\Pr(C = \text{TRUE} \mid \mathbf{i}, X_r^T = \text{TRUE}) = 1$ and thus $\text{argmax}_c \Pr(C = c, \mathbf{i}, X_r = \text{FALSE}) = \text{TRUE}$. By definition, we have that for the truth assignment $X_r = \text{FALSE}$ that $\Pr(C = \text{TRUE} \mid \mathbf{i}, X_r^T = \text{FALSE}) = 0$ and thus $\text{argmax}_c \Pr(C = c, \mathbf{i}, X_r = \text{FALSE}) = \text{FALSE}$. Hence, the intrinsic relevance $\mathcal{R}(X_r) > 0$ and thus $(\mathcal{B}_\phi, I)$ is a *yes*-instance of INTRINSIC RELEVANCE.

Now we assume that $\mathcal{R}(I) > 0$, implying that there is at least one truth assignment $\mathbf{i}$ to $\mathbf{I}\{\setminus X_r\}$ such that $\Pr(C = \text{TRUE} \mid \mathbf{i}, X_r^T = \text{FALSE}) \neq \text{argmax}_c \Pr(C = c, \mathbf{i}, X_r = \text{FALSE})$ where the joint value assignment to the operator variables O_j^T and O_j matches the deterministic conditional probabilities imposed by the joint value assignment to the variables X_i. This implies that the characteristic function $\mathbf{1}_\phi$ is not identical for $x_r = \text{FALSE}$ and $x_r = \text{TRUE}$, hence, that (ϕ, x_r) is a *yes*-instance of ISA-RELEVANT VARIABLE.

As the reduction can be done in polynomial time, this proves that INTRINSIC RELEVANCE is NP-complete. □

References

[1] Abdelbar, A.M., Hedetniemi, S.M., 1998. Approximating MAPs for belief networks is NP-hard and other theorems. Artificial Intelligence 102, 21–38.

[2] Beinlich, I., Suermondt, G., Chavez, R., Cooper, G., 1989. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks, in: Hunter, J., Cookson, J., Wyatt, J. (Eds.), Proceedings of the Second European Conference on AI and Medicine, Springer-Verlag. pp. 247–256.

[3] Bodlaender, H.L., 2006. Treewidth: characterizations, applications, and computations, in: Proceedings of the 32nd International Workshop on Graph-Theoretic Concepts in Computer Science, pp. 1–14.

[4] Bodlaender, H.L., van den Eijkhof, F., van der Gaag, L.C., 2002. On the complexity of the MPA problem in probabilistic networks, in: van Harmelen, F. (Ed.), Proceedings of the Fifteenth European Conference on Artificial Intelligence, IOS Press, Amsterdam. pp. 675–679.

[5] Bovens, L., Olsson, E.J., 2000. Coherentism, reliability and Bayesian networks. Mind 109, 686–719.

[6] Chajewska, U., Halpern, J., 1997. Defining explanation in probabilistic systems, in: Geiger, D., Shenoy, P. (Eds.), Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA. pp. 62–71.

[7] Cofiño, A.S., Cano, R., Sordo, C., Gutiérrez, J.M., 2002. Bayesian networks for probabilistic weather prediction, in: van Harmelen, F. (Ed.), Proceedings of the Fifteenth European Conference on Artificial Intelligence, IOS Press, Amsterdam. pp. 695–699.

[8] Darwiche, A., 2009. Modeling and Reasoning with Bayesian Networks. CU Press, Cambridge, UK.

[9] Darwiche, A., Choi, A., 2010. Same-decision probability: a confidence measure for threshold-based decisions under noisy sensors, in: Myllymki, P., Roos, T., Jaakkola, T. (Eds.), Proceedings of the Fifth European Workshop on Probabilistic Graphical Models, p. 113120.

[10] Demirer, R., Mau, R., Shenoy, C., 2006. Bayesian Networks: A decision tool to improve portfolio risk analysis. Journal of Applied Finance 16, 106–119.

[11] Dey, S., Stori, J.A., 2005. A Bayesian network approach to root cause diagnosis of process variations. International Journal of Machine Tools and Manufacture 45, 75–91.

[12] Downey, R.G., Fellows, M.R., 1999. Parameterized Complexity. Springer Verlag, Berlin.

[13] Druzdzel, M., 1994. Some properties of joint probability distributions, in: de Mantaras, R.L., Poole, D. (Eds.), Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, San Francisco, Ca: Morgan Kaufmann Publishers. pp. 187–194.

[14] Druzdzel, M.J., Suermondt, H.J., 1994. Relevance in probabilistic models: "backyards" in a "small world", in: Working notes of the AAAI–1994 Fall Symposium Series: Relevance, pp. 60–63.

[15] Fitelson, B., 2003. A probabilistic theory of coherence. Analysis 63, 194–199.

[16] Flum, G., Grohe, M., 2006. Parameterized Complexity Theory. Springer, Berlin.

[17] Fodor, J.A., 1983. The Modularity of Mind. Cambridge, MA: MIT Press.

[18] Fodor, J.A., 1987. Modules, frames, fridgeons, sleeping dogs, and the music of the spheres, in: Pylyshyn, Z.W. (Ed.), The Robot's Dilemma: The Frame Problem in Artificial Intelligence. Ablex Publishing, pp. 139–150.

[19] Fodor, J.A., Lepore, E., 1992. Holism: A shopper's guide. volume 16. Blackwell Oxford.

[20] Funke, J., 1991. Solving complex problems: Exploration and control of complex social systems, in: Sternberg, R.J., Frensch, P.A. (Eds.), Complex Problem Solving: Principles and Mechanisms. Lawrence Erlbaum Associates, pp. 185–222.

[21] van der Gaag, L.C., Renooij, S., Witteman, C.L.M., Aleman, B.M.P., Taal, B.G., 2002. Probabilities for a probabilistic network: a case study in oesophageal cancer. Artificial Intelligence in Medicine 25, 123–148.

[22] Garey, M.R., Johnson, D.S., 1979. Computers and Intractability. A Guide to the Theory of NP-Completeness. W. H. Freeman and Co., San Francisco, CA.

[23] Geenen, P.L., Elbers, A.R.W., van der Gaag, L.C., van der Loeffen, W.L.A., 2006. Development of a probabilistic network for clinical detection of classical swine fever, in: Proceedings of the Eleventh Symposium of the International Society for Veterinary Epidemiology and Economics, pp. 667–669.

[24] Geiger, D., Verma, T., Pearl, J., 1990. Identifying independence in Bayesian networks. Networks 20, 507–534.

[25] Gemela, J., 2001. Financial analysis using Bayesian networks. Applied Stochastic Models in Business and Industry 17, 57–67.

[26] Glass, D.H., 2007. Coherence measures and inference to the best explanation. Synthese 157, 275–296.

[27] Glass, D.H., 2009. Inference to the best explanation: a comparison of approaches, in: Bishop, M. (Ed.), Proceedings of the Second Symposium on Computing and Philosophy, The Society for the Study of Artificial Intelligence and the Simulation of Behaviour. pp. 22–27.

[28] Glass, D.H., 2012. Inference to the best explanation: does it track truth? Synthese 185, 411–427.

[29] Hempel, C.G., 1965. Aspects of Scientific Explanation. Free Press, New York.

[30] Jaynes, E., 2003. Probability Theory: The Logic of Science. Cambridge University Press.

[31] Johnson-Laird, P.N., Legrenzi, P., Girotto, V., Legrenzi, M.S., Caverni, J., 1999. Naive probability: A mental model theory of extensional reasoning. Psychological Review 106, 62–88.

[32] Kennett, R.J., Korb, K.B., Nicholson, A.E., 2001. Seabreeze prediction using Bayesian networks, in: Cheung, D.W.L., Williams, G., Li, Q. (Eds.), Proceedings of the Fifth Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Springer Verlag, Berlin. pp. 148–153.

[33] Kragt, M.E., Newhama, L.T.H., Jakemana, A.J., 2009. A Bayesian network approach to integrating economic and biophysical modelling, in: Anderssen, R., Braddock, R., Newham, L. (Eds.), Proceedings of the Eighteenth World IMACS / MODSIM Congress on Modelling and Simulation, pp. 2377–2383.

[34] Kwisthout, J., 2009. The Computational Complexity of Probabilistic Networks. Ph.D. thesis. Faculty of Science, Utrecht University, The Netherlands.

[35] Kwisthout, J., 2010. Two new notions of abduction in Bayesian networks, in: et al., P.B. (Ed.), Proceedings of the 22nd Benelux Conference on Artificial Intelligence (BNAIC'10), pp. 82–89.

[36] Kwisthout, J., 2011. Most probable explanations in Bayesian networks: Complexity and tractability. International Journal of Approximate Reasoning 52, 1452 – 1469.

[37] Kwisthout, J., 2012. Relevancy in problem solving: A computational framework. The Journal of Problem Solving 5, 17 – 32.

[38] Kwisthout, J., 2013a. Most frugal explanations: Occam's razor applied to Bayesian abduction, in: Hindriks, K., de Weerdt, M., van Riemsdijk, B., Warnier, M. (Eds.), Proceedings of the 25th Benelux Conference on AI (BNAIC'13), pp. 96–103.

[39] Kwisthout, J., 2013b. Most inforbable explanations: Finding explanations in Bayesian networks that are both probable and informative, in: van der Gaag, L. (Ed.), Proceedings of the Twelfth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, pp. 328–339.

[40] Kwisthout, J., 2013c. Structure approximation of most probable explanations in Bayesian networks, in: van der Gaag, L. (Ed.), Proceedings of the Twelfth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, pp. 340–351.

[41] Kwisthout, J., Bodlaender, H.L., van der Gaag, L.C., 2011a. The complexity of finding kth most probable explanations in probabilistic networks, in: Cerná, I., Gyimóthy, T., Hromkovic, J., Jefferey, K., Královic,

R., Vukolic, M., Wolf, S. (Eds.), Proceedings of the 37th International Conference on Current Trends in Theory and Practice of Computer Science, pp. 356–367.

[42] Kwisthout, J., van Rooij, I., 2013. Bridging the gap between theory and practice of approximate Bayesian inference. Cognitive Systems Research 24, 2–8.

[43] Kwisthout, J., Wareham, T., van Rooij, I., 2011b. Bayesian intractability is not an ailment approximation can cure. Cognitive Science 35, 779–1007.

[44] Lipton, P., 2004. Inference to the Best Explanation. London, UK: Routledge. 2nd edition.

[45] Lucas, P.J.F., de Bruijn, N., Schurink, K., Hoepelman, A., 2000. A probabilistic and decision-theoretic approach to the management of infectious disease at the ICU. Artificial Intelligence in Medicine 3, 251–279.

[46] Murphy, K., 2001. The Bayes Net Toolbox for MATLAB. Computing Science and Statistics 33, 2001.

[47] Neapolitan, R.E., 1990. Probabilistic Reasoning in Expert Systems. Theory and Algorithms. Wiley/Interscience, New York, NY.

[48] Nedevschi, S., Sandhu, J.S., Pal, J., Fonseca, R., Toyama, K., 2006. Bayesian networks: an exploratory tool for understanding ICT adoption, in: Toyama, K. (Ed.), Proceedings of the International Conference on Information and Communication Technologies and Development, pp. 277–284.

[49] Olsson, E.J., 2002. What is the problem of coherence and truth? Journal of Philosophy 99, 246–272.

[50] Papadimitriou, C.H., 1994. Computational Complexity. Addison-Wesley.

[51] Park, J.D., Darwiche, A., 2004. Complexity results and approximation settings for MAP explanations. Journal of Artificial Intelligence Research 21, 101–133.

[52] Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, Palo Alto, CA.

[53] van Rooij, I., 2008. The tractable cognition thesis. Cognitive Science 32, 939–984.

[54] Shi, L., Feldman, N., Griffiths, T., 2008. Performing Bayesian inference with exemplar models, in: Sloutsky, V., Love, B., McRae, K. (Eds.), Proceedings of the 30th annual conference of the Cognitive Science Society, pp. 745–750.

[55] Shimony, S., 1993. The role of relevance in explanation I: Irrelevance as statistical independence. International Journal of Approximate Reasoning 8, 281–324.

[56] Shimony, S.E., 1994. Finding MAPs for belief networks is NP-hard. Artificial Intelligence 68, 399–410.

[57] Stewart, N., Chater, N., Brown, G.D.A., 2006. Decision by sampling. Cognitive Psychology 53, 1–26.

[58] Sticha, P.J., Buede, D.M., Rees, R.L., 2006. Bayesian model of the effect of personality in predicting decisionmaker behavior, in: van der Gaag, L. (Ed.), Proceedings of the Fourth Bayesian Modelling Applications Workshop.

[59] Stockmeyer, L., 1977. The polynomial-time hierarchy. Theoretical Computer Science 3, 1–22.

[60] Tenenbaum, J.B., 2011. How to grow a mind: Statistics, structure, and abstraction. Science 331, 1279–1285.

[61] Torán, J., 1991. Complexity classes defined by counting quantifiers. Journal of the ACM 38, 752–773.

[62] Vul, E., Goodman, N.D., Griffiths, T.L., Tenenbaum, J.B., 2009. One and done? Optimal decisions from very few samples, in: Taatgen, N., van Rijn, H., Schomaker, L., Nerbonne, J. (Eds.), Proceedings of the 31st Annual Meeting of the Cognitive Science Society, pp. 66–72.

[63] Wagner, K.W., 1986. The complexity of combinatorial problems with succinct input representation. Acta Informatica 23, 325–356.

[64] Wasyluk, H., Onisko, A., Druzdzel, M.J., 2001. Support of diagnosis of liver disorders based on a causal Bayesian network model. Medical Science Monitor 7, 327–332.

[65] Wilson, D., Sperber, D., 2004. Relevance theory, in: R., H.L., Ward, G. (Eds.), Handbook of Pragmatics. Blackwell, Oxford, UK, pp. 607–632.

[66] Yuan, C., Lim, H., Lu, T., 2011. Most relevant explanations in Bayesian networks. Journal of Artificial Intelligence Research 42, 309–352.