Explainable AI using MAP-independence

Johan Kwisthout[1][0000−0003−4383−7786]

Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands j.kwisthout@donders.ru.nl
http://www.socsci.ru.nl/johank/

Abstract. In decision support systems the motivation and justification of the system's diagnosis or classification is crucial for the acceptance of the system by the human user. In Bayesian networks a diagnosis or classification is typically formalized as the computation of the most probable joint value assignment to the hypothesis variables, given the observed values of the evidence variables (generally known as the MAP problem). While solving the MAP problem gives the most probable explanation of the evidence, the computation is a black box as far as the human user is concerned and it does not give additional insights that allow the user to appreciate and accept the decision. For example, a user might want to know to what extent a variable was relevant for the explanation. In this paper we introduce a new concept, MAP-independence, which tries to formally capture this notion of relevance, and explore its role towards a justification of an inference to the best explanation.

Keywords: Bayesian Networks · Most Probable Explanations · Relevance · Explainable AI · Computational Complexity.

1 Introduction

With the availability of petabytes of data and the emergence of 'deep' learning as an AI technique to find statistical regularities in these large quantities of data, artificial intelligence in general and machine learning in particular has arguably entered a new phase since its emergence in the 1950s. Deep learning aims to build hierarchical models representing the data, with every new layer in the hierarchy representing ever more abstract information; for example, from individual pixels to lines and curves, to geometric patterns, to features, to categories. Superficially this might be related to how the human visual cortex interprets visual stimuli and seeks to classify a picture to be that of a cat, rather than of a dog.

When describing in what sense a cat is different from a dog, humans may use features and categories that we agreed upon to be defining features of cats and dogs, such as whiskers, location and form of the ears, the nose, etc. The deep learning method, however, does not adhere to features we humans find to be good descriptors; it bases its decisions where and how to 'carve nature's joints' solely on basis of the statistics of the data. Hence, it might very well be that the curvature of the spine (or some other apparently 'random') feature happens to be *the* statistically most important factor to distinguish cats from

dogs. This imposes huge challenges when the machine learning algorithm is asked to *motivate* its classification to a human user. The sub-field of *explainable AI* has recently emerged to study how to align statistical machine learning with informative user-based motivations or explanations. Explainable AI, however, is not limited to deep neural network applications. Any AI application where trustworthiness is important benefits from justification and transparency of its internal process [5], and this includes decision support systems that are based on Bayesian networks, which is the focus of this paper. In these systems typically one is interested in the hypothesis that best explains the available evidence; for example in a medical case, the infection that is most probable given a set of health complaints and test findings.

Note that 'explainability' in explainable AI is in principle a triadic relationship between what needs to be explained, the explanation, and the *user who seeks the explanation* [14]. An explanation will be more satisfying ('lovelier', in Peter Lipton's [10] terms) if it allows the user to gain more understanding about the phenomenon to be explained. In this paper we specifically try to improve the user's understanding of a specific decision by *explicating the relevant information* that contributed to said decision. In some way, in deciding what the best explanation is for a set of observations, the process of marginalizing out the variables that are neither observed nor hypothesis variables, makes the process more opaque: some of these variables have a bigger impact (i.e., are more relevant) on the eventual decision than others, and this information is lost in the process. For example, the absence of a specific test result (i.e., a variable we marginalize out in the MAP computation) may lead to a different explanation of the available evidence compared to when a negative (or positive) test result *were* present. In this situation, this variable is more relevant to the eventual explanation than if the best explanation would be the same, irrelevant of whether the test result was positive, negative, or missing. Our approach in this paper is to motivate a decision by showing which of these variables were relevant in this sense towards arriving at this decision.

This perspective has roots in Pearl's early work on conditional independence [13]. Pearl suggests that human reasoning is in principle based on *conditional independence*: The organizational structure of human memory is such that it allows for easily retrieving context-dependent relevant information. For example (from [13, p.3]): The color of my friend's car is normally not related to the color of my neighbour's car. However, when my friend tells me she almost mistook my neighbour's car from her own, this information suddenly becomes relevant for me to understand, and for her to explain, this mistake. That is, the color of both cars is independent but becomes conditionally dependent on the evidence[1].

[1] Graphically one can see this as a so-called common-effect structure, where C_1 and C_2 are variables that represent my car's, respectively my neighbour's car's, color; both variables have a directed edge towards the variable M that indicates whether my friend misidentified the cars or not. When M is unobserved, C_1 and C_2 are independent, but they become conditionally dependent on observation of M.

In this paper we will argue that Pearl's proposal to model context-dependent (ir)relevance as conditional (in)dependence is in fact too strict. It generally leads to too many variables that are considered to be relevant: for some it is likely the case that, while they may not be conditionally independent on the hypothesized explanation given the evidence, they do not contribute to understanding *why* some explanation h is better than the alternatives. That means, for explanatory purposes their role is limited. In the remainder of this paper we will build on Pearl's work, yet provide a stronger notion of context-dependent relevance and irrelevance of variables relative to explanations of observations. Our goal is to further explainable AI in the context of Bayesian networks by formalizing the problem of *justification* of an explanation (i.e., given an AI-generated explanation, advance the user's understanding why this explanation is preferred over others) into a computational problem that captures some aspects of this justification; in particular, by opening up the 'marginalization black box' and show which variables contributed to this decision. We show that this problem is intractable in the general case, but also give fixed-parameter tractability results that show what constraints are needed to render it tractable.

To summarize, we are interested in the potential applicability of this new concept for motivation and justification of MAP explanations, with a focus on its theoretical properties. The remainder of this paper is structured as follows. In the next section we offer some preliminary background on Bayesian networks and computational complexity and share our conventions with respect to notation with the reader. In section 3 we introduce so-called MAP-independence as an alternative to conditional independence and elaborate on the potential of these computational problems for justifying explanations in Bayesian networks. In section 4 we introduce a formal computational problem based on this notion, and give complexity proofs and fixed-parameter tractability results for this problem. We conclude in section 5.

2 Preliminaries and notation

In this section we give some preliminaries and introduce the notational conventions we use throughout this paper. The reader is referred to textbooks like [1] for more background.

A Bayesian network $\mathcal{B} = (\mathbf{G}_\mathcal{B}, \Pr)$ is a probabilistic graphical model that succinctly represents a joint probability distribution $\Pr(\mathbf{V}) = \prod_{i=1}^{n} \Pr(V_i \mid \pi(V_i))$ over a set of discrete random variables $\mathbf{V}$. $\mathcal{B}$ is defined by a directed acyclic graph $\mathbf{G}_\mathcal{B} = (\mathbf{V}, \mathbf{A})$, where $\mathbf{V}$ represents the stochastic variables and $\mathbf{A}$ models the conditional (in)dependencies between them, and a set of parameter probabilities $\Pr$ in the form of conditional probability tables (CPTs). In our notation $\pi(V_i)$ denotes the set of parents of a node V_i in $\mathbf{G}_\mathcal{B}$. We use upper case to indicate variables, lower case to indicate a specific value of a variable, and boldface to indicate sets of variables respectively joint value assignments to such a set. $\Omega(V_i)$ denotes the set of value assignments to V_i, with $\Omega(\mathbf{V_a})$ denoting the set of joint value assignment to the set $\mathbf{V_a}$.

One of the key computational problems in Bayesian networks is the problem to find the most probable explanation for a set of observations, i.e., a joint value assignment to a designated set of variables (the explanation set) that has maximum posterior probability given the observed variables (the joint value assignment to the evidence set) in the network. If the network includes variables that are neither observed nor to be explained (referred to as intermediate variables) this problem is typically referred to as MAP. We use the following formal definition:

MAP
Instance: A Bayesian network $\mathcal{B} = (\mathbf{G}_\mathcal{B}, \Pr)$, where $\mathbf{V}(\mathbf{G}_\mathcal{B})$ is partitioned into a set of evidence nodes $\mathbf{E}$ with a joint value assignment $\mathbf{e}$, a set of intermediate nodes $\mathbf{I}$, and an explanation set $\mathbf{H}$.
Output: A joint value assignment $\mathbf{h}^*$ to $\mathbf{H}$ such that for all joint value assignments $\mathbf{h}'$ to $\mathbf{H}$, $\Pr(\mathbf{h}^* \mid \mathbf{e}) \geq \Pr(\mathbf{h}' \mid \mathbf{e})$.

We assume that the reader is familiar with standard notions in computational complexity theory, notably the classes P and NP, NP-hardness, and polynomial time (many-one) reductions. The class PP is the class of decision problems that can be decided by a probabilistic Turing machine in polynomial time; that is, where *yes*-instances are accepted with probability strictly larger than $1/2$ and *no*-instances are accepted with probability no more than $1/2$. A problem in PP might be accepted with probability $1/2 + \epsilon$ where ϵ may depend exponentially on the input size n. Hence, it may take exponential time to increase the probability of acceptance (by repetition of the computation and taking a majority decision) close to 1. PP is a powerful class; we know for example that $\mathsf{NP} \subseteq \mathsf{PP}$ and the inclusion is assumed to be strict. The canonical PP-complete decision problem is MAJSAT: given a Boolean formula ϕ, does the majority of truth assignments to its variables satisfy ϕ?

In computational complexity theory, so-called *oracles* are theoretical constructs that increase the power of a specific Turing machine. An oracle (e.g., an oracle for PP-complete problems) can be seen as a 'magic sub-routine' that answers class membership queries (e.g, in PP) in a single time step. In this paper we are specifically interested in classes defined by non-deterministic Turing machines with access to a PP-oracle. Such a machine is very powerful, and likewise problems that are complete for the corresponding complexity classes $\mathsf{NP}^{\mathsf{PP}}$ (such as MAP) and co-$\mathsf{NP}^{\mathsf{PP}}$ (such as MONOTONICITY) are highly intractable [12, 3].

3 MAP-independence

The topic of *relevance* in Bayesian networks has been studied from several angles: while [17] aimed to reduce the number of variables in the explanation set to the relevant ones, and [11] studied the relevance of evidence variables for sensitivity analysis, in [9] the approach was to reduce the number of intermediate variables that affect an inference to the best explanation. In the current paper we take a similar approach, but here we focus on the application of this notion of relevance

in explainable AI, rather than to construct a heuristic approach towards the computationally expensive MAP problem. Here the problem of interest is not so much to *find* the most probable explanation (viz., the joint value assignment to a set of hypothesis variables given observations in the network), but rather to *motivate* what information did or did not contribute to a given explanation.

That is, rather than providing the 'trivial' explanation "$\mathbf{h}^*$ is the best[2] explanation for $\mathbf{e}$, since $\text{argmax}_{\mathbf{h}} \Pr(\mathbf{H} = \mathbf{h} \mid \mathbf{e}) = \text{argmax}_{\mathbf{h}} \sum_{\mathbf{i} \in \Omega(\mathbf{I})} \Pr(\mathbf{H} = \mathbf{h}, \mathbf{i} \mid \mathbf{e}) = \mathbf{h}^*$" our goal is to partition the set $\mathbf{I}$ into variables I^+ that are *relevant* to establishing the best explanation and variables I^- that are *irrelevant*. One straightforward approach, motivated by [13], would be to include variables in I^+ if they are conditionally dependent on $\mathbf{H}$ given $\mathbf{e}$, and in I^- when they are conditionally independent from $\mathbf{H}$ given $\mathbf{e}$ and to *motivate* the sets I^+ and I^- in terms of a set of independence relations. This is particularly useful when inclusion in I^+ is *triggered* by the presence of an observation, such as in Pearl's example where 'color of my friend's car' and 'color of my neighbour's car' become dependent on each other once we learn that my friend confused both cars.

We argue that this way of partitioning intermediate variables into relevant and irrelevant ones, however useful, might not be the full story with respect to explanation. There is a sense in which a variable has an explanatory role in motivating a conclusion that goes beyond conditional (in)dependence. Take for example the small binary network in Figure 1. Assume that we want to motivate the best explanation for A given the evidence $C = c$, i.e., we want to motivate the outcome of $\text{argmax}_a \Pr(A = a \mid C = c) = \text{argmax}_a \sum_{B,D} \Pr(A = a, B, D \mid C = c)$ in terms of variables that contribute to this explanation. Now, obviously D is *not* relevant, as it is d-separated from A given C. But the roles of B is less obvious. This node is obviously not conditionally independent from A given C.

Whether B plays an explanatory role in general in the outcome of the MAP query is dependent on whether $\text{argmax}_a \sum_D \Pr(A = a, B = b, D \mid C = c) = \text{argmax}_a \sum_D \Pr(A = a, B = \bar{b}, D \mid C = c)$. If both are equal ($B$'s value, were it observed, would have been irrelevant to the MAP query) than B arguably has no explanatory role. If both are unequal than the fact that B is unobserved may in fact be crucial for the explanation. For example, if B represents a variable that encodes a 'default' versus 'fault' condition (with $\Pr(b) > \Pr(\bar{b})$) the *absence* of a fault (i.e., B is unobserved) can lead to a different MAP explanation than the *observation* that B takes its default value; e.g., if $\Pr(b)) = 0.6$, $\Pr(a \mid c, b) = 0.6$, and $\Pr(a \mid c, \bar{b}) = 0.3$ we have that $\Pr(a \mid c) = 0.48$ yet $\Pr(a \mid c, b) = 0.6$ so the MAP explanation changes from $\bar{a}$ to a on the observation of the *default* value b. This information is lost in the marginalization step but it helps motivate *why* $\bar{a}$ is the best explanation of d.

[2] In this paper we do not touch the question whether 'best' is to be identified with 'most probable'. The interested reader is referred to the vast literature on inference to the best explanation such as [10], and more in particular to some of our earlier work [8] that discusses the trade-off between probability and informativeness of explanations.

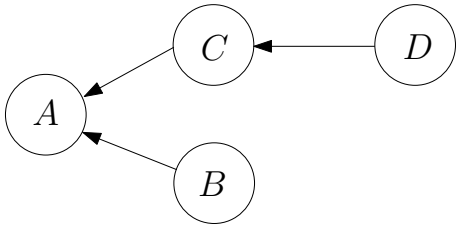

Fig. 1. An example small network. Note that the explanatory role of B in motivating the best explanation of A given an observation for C is context-dependent and may be different for different observations for C, as well as for different conditional probability distributions; hence it cannot be read off the graph alone.

Thus, the relevance of B for the explanation of A may need a *different* (and broader) notion of independence, as also suggested by [9]. Indeed, in this example, variable D is irrelevant for explaining A as D is conditionally independent from A given C; yet, we could also argue that B is irrelevant for explaining A *if its value, were it observed, could not influence the explanation for* A. We introduce the term *MAP-independence* for this (uni-directional) relationship; we say that A is MAP-independent from B given $C = c$ when $\forall_{b \in \Omega(B)} \text{argmax}_a \Pr(A = a, B = b \mid C = c) = a$ for a specific value assignment $a \in \Omega(A)$.

3.1 MAP-independence for justification and decision support

AI-based clinical decision support systems for diagnosis and treatment have been proposed since the 1970s, with MYCIN [15] as the canonical example. Whereas original systems were largely an effort to demonstrate the promise of AI techniques (i.e., isolated, difficult to maintain or generalize, of mostly academic interest, etc.), current systems have developed into systems that are integrated with the medical workflow, in particular aligned with electronic health records [16]. However, several challenges that were already identified with MYCIN still remain present in decision support systems: they are difficult to maintain and adapt to new insights, the justification of the system's advice does not match typical reasoning patterns of the user, and there is little justification of the soundness (or acknowledgement of uncertainty and ignorance) of the advice.

The concept of MAP-independence in Bayesian networks may help overcome some of these shortcomings, particularly the justification of an inference to the most probable explanation. For an unobserved variable I we have that the MAP explanation $\mathbf{h}*$ is MAP-dependent of I given the evidence if the explanation would not have been different had I be observed to some value, and MAP-independent if this is not the case. An explication of how I may impact or fail to impact the most probable explanation of the evidence will both help motivate the system's advice as well as offer guidance in further decisions (e.g., to gather additional evidence [4, 2] to make the MAP explanation more robust).

4 Formal problem definition and results

The computational problem of interest is to decide upon the set I^+, the relevant variables that contribute to establishing the best explanation $\mathbf{h}^*$ given the evidence $\mathbf{e}$. In order to establish I^+ we need to decide the following sub-problem: given $\mathbf{R} \subseteq \mathbf{I}$: is $\mathbf{H}$ MAP-independent from $\mathbf{R}$ given $\mathbf{e}$? We formalize this problem as below.

MAP-INDEPENDENCE
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 $\mathbf{e}$, a non-empty explanation set $\mathbf{H}$ with a joint value assignment $\mathbf{h}^* = \text{argmax}_\mathbf{h} \Pr(\mathbf{H} = \mathbf{h}, \mathbf{e})$, a non-empty set of nodes $\mathbf{R}$ for which we want to decide MAP-independence relative to $\mathbf{H}$, and a set of intermediate nodes $\mathbf{I}$.
Question: Is $\forall_{\mathbf{r} \in \Omega(\mathbf{R})} \text{argmax}_\mathbf{H} \Pr(\mathbf{H}, \mathbf{R} = \mathbf{r} \mid \mathbf{e}) = \mathbf{h}^*$?

Observe that the complement problem MAP-DEPENDENCE is defined similarly with *yes*- and *no*-answers reversed.

4.1 Computational complexity

We will show in this sub-section that an appropriate decision variant of MAP-INDEPENDENCE is co-$\mathsf{NP}^{\mathsf{PP}}$-complete and thus resides in the same complexity class as the MONOTONICITY problem [3]. Note that the definition of MAP-INDEPENDENCE in the previous sub-section had the MAP explanation given in the input and assumed that $\mathbf{R}$ is nonempty. The reason therefore is that if we would allow $\mathbf{R} = \varnothing$ and leave out the MAP explanation, the problem has MAP as a degenerate special case. As this somewhat obfuscates the computational complexity of the core of the problem (i.e., determining the relevance of the set $\mathbf{R}$ for the actual explanation) we force $\mathbf{R}$ to be non-empty and $\text{argmax}_\mathbf{h} \Pr(\mathbf{H} = \mathbf{h}, \mathbf{e})$ to be provided in the input.

Another complication is that, while MAP-INDEPENDENCE is already defined as a decision problem, part of the problem definition requires comparing MAP explanations, and while the MAP problem has a decision variant that is $\mathsf{NP}^{\mathsf{PP}}$-complete, the functional variant is $\mathsf{FP}^{\mathsf{NP}^{\mathsf{PP}}}$-complete [6]. We therefore introduce the following decision variant[3] which is in the line with the traditional decision variant of PARTIAL MAP:

[3] Note that as a decision variant of MAP-INDEPENDENCE there is still a slight caveat, as the probability of $\Pr(\mathbf{h}^*, \mathbf{r}, \mathbf{e})$ can be different for each joint value assignment $\mathbf{r}$, implying that the 'generic' threshold s can either be too strict ($\mathbf{h}$ is still the MAP explanation although the test fails) or too loose (there is another explanation $\mathbf{h}'$ which is the MAP explanation although the test passes). As the number of joint value assignments $|\mathbf{r}|$ can be exponential in the size of the network (and thus we cannot include individual thresholds s_i in the input of the decision problem without blowing up the input size), this nonetheless appears to be the closest decision problem variant that still captures the crucial aspects of MAP-INDEPENDENCE.

MAP-INDEPENDENCE-D
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 $\mathbf{e}$, a non-empty explanation set $\mathbf{H}$ with a joint value assignment $\mathbf{h}^* = \text{argmax}_\mathbf{h} \Pr(\mathbf{H} = \mathbf{h}, \mathbf{e})$, a non-empty set of nodes $\mathbf{R}$ for which we want to decide MAP-independence relative to $\mathbf{H}$, and a set of intermediate nodes $\mathbf{I}$; rational number s.
Question: Is, for each joint value assignment $\mathbf{r}$ to $\mathbf{R}$, $\Pr(\mathbf{h}^*, \mathbf{r}, \mathbf{e}) > s$?

For the hardness proof we reduce from the canonical satisfiability co-$\mathsf{NP}^{\mathsf{PP}}$-complete variant A-MAJSAT defined as follows:

A-MAJSAT
Instance: A Boolean formula ϕ with n variables $\{x_1, x_n\}$, partitioned into the sets $\mathbf{A} = \{x_1, x_k\}$ and $\mathbf{M} = \{x_{k+1}, x_n\}$ for some $k \leq n$.
Question: Does, for every truth instantiation $\mathbf{x_a}$ to $\mathbf{A}$, the majority of truth instantiations $\mathbf{x_m}$ to $\mathbf{M}$ satisfy ϕ?

As a running example for our reduction we use the formula $\phi_{ex} = \neg(x_1 \wedge x_2) \vee (x_3 \vee x4)$, with $\mathbf{A} = \{x_1, x_2\}$ and $\mathbf{M} = \{x_3, x_4\}$. This is a *yes*-instance of A-MAJSAT: for each truth assignment to $\mathbf{A}$, at least three out of four truth assignments to $\mathbf{M}$ satisfy ϕ.

We construct a Bayesian network $\mathcal{B}_\phi$ from a given Boolean formula ϕ with n variables. For each propositional variable x_i in ϕ, a binary stochastic variable X_i is added to $\mathcal{B}_\phi$, with possible values T and F and a uniform probability distribution. 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. For example, the value T of a stochastic variable mimicking the *and*-operator would have a conditional probability of 1 if and only if both its parents have the value T, and 0 otherwise. The top-level operator in ϕ is denoted as V_ϕ. In Figure 2 the network $\mathcal{B}_\phi$ is shown for the formula $\neg(x_1 \wedge x_2) \vee (x_3 \vee x4)$.

Theorem 1. MAP-INDEPENDENCE *is co-*$\mathsf{NP}^{\mathsf{PP}}$*-complete.*

Proof. To prove membership in co-$\mathsf{NP}^{\mathsf{PP}}$, we give a falsification algorithm for *no*-answers to MAP-INDEPENDENCE-D instances, given access to an oracle for the PP-complete INFERENCE problem. Let $(\mathcal{B}, \mathbf{E}, \mathbf{e}, \mathbf{H}, \mathbf{h}^*, \mathbf{R}, \mathbf{I}, s)$ be an instance of MAP-INDEPENDENCE-D. We non-deterministically guess a joint value assignment $\bar{\mathbf{r}}$, and use the INFERENCE oracle to verify that $\Pr(\mathbf{h}^*, \bar{\mathbf{r}}, \mathbf{e}) \leq s$, which by definition implies that $\mathbf{H}$ is not MAP-independent from $\mathbf{R}$ given $\mathbf{E} = \mathbf{e}$.

To prove hardness, we reduce from A-MAJSAT. Let $(\phi, \mathbf{X_A}, \mathbf{X_M})$ be an instance of A-MAJSAT and let $\mathcal{B}_\phi$ be the Bayesian network created from ϕ as per the procedure described above. We set $\mathbf{R} = \mathbf{X_A}$, $\mathbf{I} = \mathbf{X_M}$, $\mathbf{H} = \{V_\phi\}$, $\mathbf{h}^* = \{T\}$, $\mathbf{E} = \varnothing$, and $s = 2^{-|\mathbf{R}|-1}$.

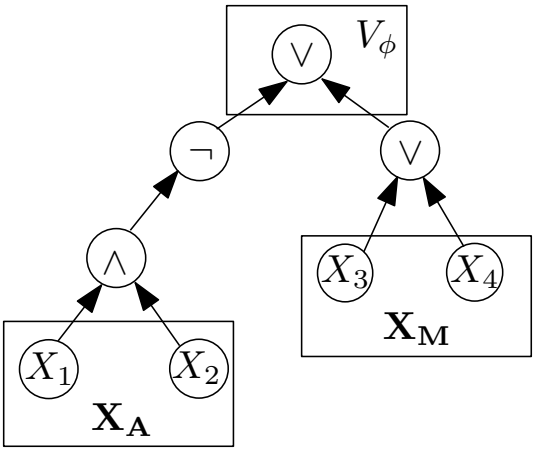

Fig. 2. The network $\mathcal{B}_\phi$ created from the A-MAJSAT instance $(\phi, \{x_1, x_2\}, \{x_3, x_4\})$ per the description above.

$\Longrightarrow$ Assume that $(\phi, \mathbf{X_A}, \mathbf{X_M})$ is a *yes*-instance of A-MAJSAT, i.e., for every truth assignment to $\mathbf{X_A}$, the majority of truth assignments to $\mathbf{X_M}$ satisfies ϕ. Then, by the construction of $\mathcal{B}_\phi$, we have $\sum_{\mathbf{r}} \Pr(V_\phi = T, \mathbf{r}) > 1/2$ and so, as the variables in $\mathbf{R}$ are all uniformly distributed, $\Pr(V_\phi = T, \mathbf{r}) > 2^{-|\mathbf{R}|-1}$ for every joint value assignment $\mathbf{r}$ to $\mathbf{R}$, and so this is a *yes*-instance of MAP-INDEPENDENCE-D.

$\Longleftarrow$ Assume that $(\mathcal{B}, \varnothing, \varnothing, V_\phi, T, \mathbf{R}, \mathbf{I}, 2^{-|\mathbf{R}|-1})$ is a *yes*-instance of MAP-INDEPENDENCE-D. Given the construction this implies that for all joint value assignments $\mathbf{r}$ it holds that $\Pr(V_\phi = T, \mathbf{r}) > 2^{-|\mathbf{R}|-1}$. But this implies that for all truth assignments to $\mathbf{X_A}$, the majority of truth assignments to $\mathbf{X_M}$ satisfies ϕ, hence, this is a *yes*-instance of A-MAJSAT.

Observe that the construction of $\mathcal{B}_\phi$ takes time, polynomial in the size of ϕ, which concludes our proof. Furthermore, the results holds in the absence of evidence

Corollary 1. MAP-DEPENDENCE *is* $\mathsf{NP}^{\mathsf{PP}}$*-complete.*

4.2 Algorithm and algorithmic complexity

To decide whether a MAP explanation $\mathbf{h}^*$ is MAP-independent from a set of variables $\mathbf{R}$ given evidence $\mathbf{e}$, the straightforward algorithm below shows that the run-time of this algorithm is $\mathcal{O}(\Omega(\mathbf{R})) = \mathcal{O}(2^{|\mathbf{R}|})$ times the time needed for each MAP computation.

Algorithm 1: Straightforward MAP-INDEPENDENCE algorithm

Input: Bayesian network partitioned in $\mathbf{E} = \mathbf{e}$, $\mathbf{H} = \mathbf{h}^*$, $\mathbf{R}$, and $\mathbf{I}$.
Output: *yes* if $\mathbf{H}$ is MAP-independent from $\mathbf{R}$ given $\mathbf{e}$, *no* if otherwise.
foreach $\mathbf{r} \in \Omega(\mathbf{R})$ **do**
  if $\text{argmax}_{\mathbf{H}} \Pr(\mathbf{H}, \mathbf{R} = \mathbf{r} \mid \mathbf{e}) \neq \mathbf{h}^*$ **then**
    return no;
  end
end
return yes;

This implies that, given known results on fixed-parameter tractability [7] and efficient approximation[12, 9] of MAP, the size of the set against which which we want to establish MAP independence is the crucial source of complexity if MAP can be computed or approximated feasibly. The following fixed-parameter tractability results can be derived:

Corollary 2. *Let* $c = \max_{W \in V(G)} \Omega(W)$, $q = \Pr(\mathbf{h}^*)$, *and let* tw *be the treewidth of* $\mathcal{B}$. *Then* p-MAP-INDEPENDENCE *is fixed-parameter tractable for* $p = \{|\mathbf{H}|, |\mathbf{R}|, \text{tw}, c\}$, $p = \{|\mathbf{H}|, |\mathbf{R}|, |\mathbf{I}|, c\}$, $p = \{q, |\mathbf{R}|, \text{tw}, c\}$, *and* $p = \{q, |\mathbf{R}|, |\mathbf{I}|, c\}$.

5 Conclusion and future work

In this paper we introduced MAP-independence as a formal notion, relevant for decision support and justification of decisions. In a sense, MAP-independence is a relaxation of conditional independence, suggested by Pearl [13] to be a scaffold for human context-dependent reasoning. We suggest that that MAP-independence may be a useful notion to further explicate the variables that are *relevant* for the establishment of a particular MAP explanation. Establishing whether the MAP explanation is MAP-independent from a set of variables given the evidence (and so, whether these variables are relevant for justifying the MAP explanation) is a computationally intractable problem; however, for a *specific* variable of interest I (or a small set of these variables together) the problem is tractable whenever MAP can be computed tractably; in practice, this may suffice for usability in typical decision support systems.

There are many related problems of interest that one can identify, but which will be delegated to future work. For example, if the set of relevant variables is large, one might be interested in deciding whether observing one variable can bring down this set (by more than one, obviously). Another related problem would be to decide upon the observations that are relevant for the MAP explanation (i.e., had we not observed $E \in \mathbf{E}$ or had we observed a different value, would that change the MAP explanation?) This would extend previous work [11] where the relevance of E for computing a posterior probability (conditioned on $\mathbf{E}$) was established. Finally, in order to test its practical usage, the formal concept introduced in this paper should be put to the empirical test in an actual decision support system to establish whether the justifications supported

by the notion of MAP-independence actually help understand and appreciate the system's advise.

References

1. Darwiche, A.: Modeling and Reasoning with Bayesian Networks. Cambridge University Press (2009)
2. van der Gaag, L., Bodlaender, H.: On stopping evidence gathering for diagnostic Bayesian networks. In: European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty. pp. 170–181 (2011)
3. van der Gaag, L., Bodlaender, H., Feelders, A.: Monotonicity in Bayesian networks. In: Chickering, M., Halpern, J. (eds.) Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence. pp. 569–576. Arlington: AUAI press (2004)
4. van der Gaag, L., Wessels, M.: Selective evidence gathering for diagnostic belief networks. AISB Quarterly **86**, 23–34 (1993)
5. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.: XAI—explainable artificial intelligence. Science Robotics **4**(37) (2019)
6. Kwisthout, J.: Complexity results for enumerating MPE and Partial MAP. In: Jaeger, M., Nielsen, T. (eds.) Proceedings of the Fourth European Workshop on Probabilistic Graphical Models. pp. 161–168 (2008)
7. Kwisthout, J.: Most Probable Explanations in Bayesian networks: Complexity and tractability. International Journal of Approximate Reasoning **52**(9) (2011)
8. Kwisthout, J.: 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. LNAI, vol. 7958, pp. 328–339. Springer-Verlag (2013)
9. Kwisthout, J.: Tree-width and the computational complexity of MAP approximations in Bayesian networks. Journal of Artificial Intelligence Research **53**, 699–720 (2015)
10. Lipton, P.: Inference to the Best Explanation. London, UK: Routledge (1991)
11. Meekes, M., Renooij, S., van der Gaag, L.: Relevance of evidence in Bayesian networks. In: European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty. pp. 366–375. Springer (2015)
12. Park, J.D., Darwiche, A.: Complexity results and approximation settings for MAP explanations. Journal of Artificial Intelligence Research **21**, 101–133 (2004)
13. Pearl, J., Paz, A.: GRAPHOIDS: a graph-based logic for reasoning about relevance relations. Tech. Rep. R-53-L, UCLA Computer Science Department (1987)
14. Ras, G., van Gerven, M., Haselager, W.: Explanation methods in deep learning: Users, values, concerns and challenges. In: Explainable and Interpretable Models in Computer Vision and Machine Learning, pp. 19–36. Springer (2018)
15. Shortliffe, E., Buchanan, B.: A model of inexact reasoning in medicine. Mathematical Biosciences **379**, 233–262 (1975)
16. Sutton, R., Pincock, D., Baumgart, D., Sadowski, D., Fedorak, R., Kroeker, K.: An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ digital medicine **3**(1), 1–10 (2020)
17. Yuan, C., Lim, H., Lu, T.: Most relevant explanation in Bayesian networks. Journal of Artificial Intelligence Research **42**(1), 309–352 (2011)