Relevance for Robust Bayesian Network MAP-Explanations

Silja Renooij s.renooij@uu.nl
Department of Information and Computing Sciences, Utrecht University, The Netherlands

Abstract

In the context of explainable AI, the concept of MAP-independence was recently introduced as a means for conveying the (ir)relevance of intermediate nodes for MAP computations in Bayesian networks. In this paper, we further study the concept of MAP-independence, discuss methods for finding sets of relevant nodes, and suggest ways to use these in providing users with an explanation concerning the robustness of the MAP result.

Keywords: Bayesian networks; relevance; explainability; MAP-independence; robustness.

1. Introduction

Among the methods for explaining Bayesian networks are those for explaining the evidence, such as the MAP-explanation (Lacave and Díez, 2002). A MAP-explanation is the outcome of computing the maximum a-posteriori (MAP) hypothesis, i.e. the most probable hypothesis given the observed evidence. In answering a MAP query, all nodes that are neither hypothesis nodes nor evidence nodes are summed out. These intermediate nodes, however, can carry important information about the stability of the MAP-explanation: will the MAP-explanation be the same regardless of the values of the intermediate nodes? To capture the (ir)relevance of the intermediate nodes to this end, the concept of MAP-*independence* was introduced as a tool that can be used to improve a user's understanding of the MAP-explanation by revealing the information that did or did not contribute to it (Kwisthout, 2021). In Bayesian network literature, the word *relevance* has been used for various purposes, to denote different properties of different sets of nodes (see Kwisthout (2021) for a discussion of this related work). Unlike most of this related literature, the notion of relevance tied to the concept of MAP-independence pertains to intermediate nodes only and is motivated by its potential application in explainable AI, for explaining a MAP-explanation's stability.

In addition to motivating and introducing the concept of MAP-independence, Kwisthout (2021) studies the computational complexity of determining whether a MAP-explanation is MAP-independent of a *given* subset $\mathbf{R}$ of intermediate nodes; this decision problem turns out to be co-NP$^{\text{PP}}$-complete. Valero-Leal et al. (2022) provide properties that can be exploited to improve the computational efficiency of determining MAP-independence and formulate a similar concept for networks with continuous variables. Our main focus in this paper is on determining *which* subsets $\mathbf{R}$ are of interest to investigate in terms of MAP-independence. In doing so we study theoretical properties of MAP-independence and consider how to employ the obtained results for explaining the robustness of a MAP-explanation to future observations for intermediate nodes. For the latter purpose, we consider to what extent we can meet the criteria that are important for explainable AI: explanations are contrastive, selected, do not refer to probabilities and are social (Miller, 2019).

Bayesian network classifiers typically return a MAP-explanation as output; in case of a single class node, the output can be considered a special case of a MAP query for a single hypothesis node. Explanation methods for Bayesian network classifiers mostly focus on directly relating input (evidence) and output (see e.g. Koopman and Renooij (2021) for a brief overview). Recognizing the importance of including intermediate variables in the explanation of Bayesian network classifiers, *influence driven* explanations were introduced that include most likely values of intermediate nodes for Bayesian networks with a constrained structure (Albini et al., 2021). MAP-independence provides an alternative for including intermediate nodes in the explanation, without making any structural assumptions.

This paper is organised as follows. We present some relevant preliminaries, terminology and notation in Section 2. Section 3 defines (ir)relevance and its properties. Finding relevant sets is the topic of Section 4 and their use for explaining robustness of MAP-explanations is discussed in Section 5. The paper is concluded in Section 6.

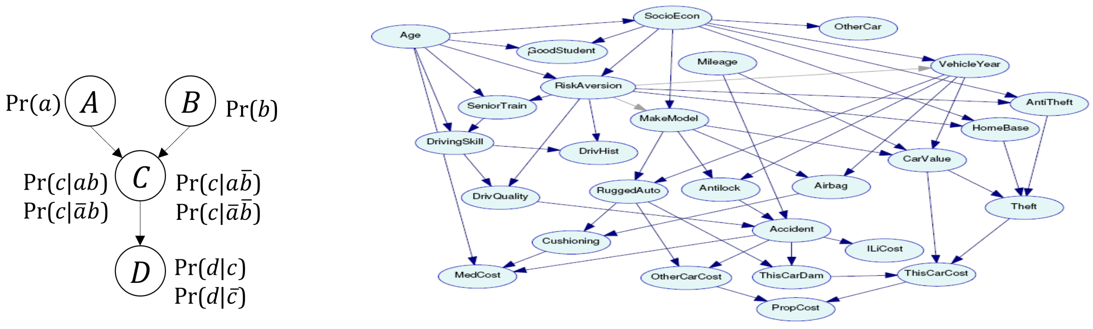
2. Preliminaries

We consider a Bayesian network (BN) $\mathcal{B} = (G, \Pr)$ representing a joint probability distribution $\Pr(\mathbf{V})$ over a set of discrete random variables $\mathbf{V}$ (Jensen and Nielsen, 2007). We use capital letters to denote variables, bold-faced in case of sets. We use $\Omega(V)$ to represent the domain of variable $V \in \mathbf{V}$, writing v as shorthand for a value assignment $V = v$, $v \in \Omega(V)$; for binary-valued variables we use v and $\overline{v}$. Bold-faced small letters $\mathbf{v}$ denote joint value assignments, or *configurations* from $\Omega(\mathbf{V})$; the latter captures the domain of all configurations of $\mathbf{V}$. Each variable is represented by a node in the directed acyclic graph G, which captures the independence relation among $\mathbf{V}$ through the notion of *d-separation* (Pearl, 1988). With each variable, or node, V is associated a set of local probability distributions $\Pr(V \mid \boldsymbol{\pi}(V))$, one for each configuration of parents $\boldsymbol{\pi}(V)$ of V in the graph, that together define the joint distribution: $\Pr(\mathbf{V}) = \prod_{V \in \mathbf{V}} \Pr(V \mid \boldsymbol{\pi}(V))$. Figure 1a shows an example Bayesian network with four binary-valued nodes and the required probability parameters.

In this paper we are interested in explaining the relevance of intermediate nodes with respect to an outcome of interest. To this end we assume that $\mathbf{V}$ is partitioned into three disjoint sets: hypothesis nodes ($\mathbf{H}$), evidence nodes ($\mathbf{E}$) and the remaining nodes $\mathbf{S} = \mathbf{V} \setminus (\mathbf{H} \cup \mathbf{E})$. Although the nodes in $\mathbf{S}$ are often referred to as 'intermediate nodes' or 'hidden nodes', we will call them *supplementary nodes*. The reason for this is that the term 'intermediate' may suggest that these nodes are on a chain between $\mathbf{H}$ and $\mathbf{E}$, which is not necessarily the case; the term 'hidden' often entails that the nodes are not observable, whereas we assume that $\mathbf{S}$ can include observable nodes that currently have no evidence. The set $\mathbf{E}$, therefore, contains only the currently observed nodes and not necessarily all observable ones. Our outcome of interest is the MAP-explanation, i.e. the most likely hypothesis $\mathbf{h}^* \in \Omega(\mathbf{H})$ to explain the evidence $\mathbf{e} \in \Omega(\mathbf{E})$:

$$\mathbf{h}^* = \arg\max_{\mathbf{h}' \in \Omega(\mathbf{H})} \Pr(\mathbf{h}' \mid \mathbf{e}) = \arg\max_{\mathbf{h}' \in \Omega(\mathbf{H})} \Pr(\mathbf{h}'\,\mathbf{e}) = \arg\max_{\mathbf{h}' \in \Omega(\mathbf{H})} \sum_{\mathbf{s} \in \Omega(\mathbf{S})} \Pr(\mathbf{h}'\,\mathbf{s}\,\mathbf{e})$$

Note that the supplementary nodes are the nodes summed out in the computation of the MAP-explanation. We will refer to the tuple $\langle \mathbf{h}^*, \mathbf{e} \rangle$ as the *explanation context*. Since we



(a) (b)

Figure 1: (a) a simple BN, (b) the graph of the Insurance BN (Binder et al., 1997).

do not assume the Bayesian network to be a causal model, the MAP-explanation is *not* necessarily a *causal* explanation for the observed evidence (Halpern and Pearl, 2005).

3. Relevance: Definition and Properties

For the purpose of explaining the actual contribution of the supplementary nodes in computing the MAP-explanation, Kwisthout (2021) proposes to partition the set $\mathbf{S}$ of supplementary nodes into a set $\mathbf{S}^+$ of relevant nodes and a set $\mathbf{S}^-$ of irrelevant nodes. Where Pearl and Paz (1987) define irrelevance in terms of d-separation, Kwisthout (2021) argues that "there is a sense in which a variable has an explanatory role [...] that goes beyond conditional (in)dependence". Kwisthout (2021) therefore captures the irrelevance of supplementary nodes in terms of a newly defined concept of MAP-independence.

Definition 1 (Kwisthout (2021)) *Let* $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ *be as before. Let* $\mathbf{R} \subseteq \mathbf{S}$ *and consider evidence context* $\langle \mathbf{h}^*, \mathbf{e} \rangle$*. In the given context,* $\mathbf{h}^*$ *is* MAP-independent *of* $\mathbf{R}$ *if for all* $\mathbf{r} \in \Omega(\mathbf{R})$*,* $\arg\max_{\mathbf{h}' \in \Omega(\mathbf{H})} \Pr(\mathbf{h}'\,\mathbf{r} \mid \mathbf{e}) = \mathbf{h}^*$*.*

If a MAP-explanation is MAP-independent of a set of unobserved nodes $\mathbf{R}$, the most likely hypothesis cannot change upon obtaining evidence for $\mathbf{R}$, in which case $\mathbf{R}$ is considered irrelevant to the explanation. To capture this, we explicitly define (ir)relevance as follows.

Definition 2 *Let* $\mathbf{H}, \mathbf{E}, \mathbf{S}, \mathbf{R}$*, and* $\langle \mathbf{h}^*, \mathbf{e} \rangle$ *be as before. In the given context,* $\mathbf{R}$ *is* relevant *to* $\mathbf{h}^*$ *if* $\mathbf{h}^*$ ***is not*** MAP*-independent of* $\mathbf{R}$*; otherwise* $\mathbf{R}$ *is* irrelevant *to* $\mathbf{h}^*$*.*

Kwisthout (2021) focuses on the computational complexity of determining whether a given set $\mathbf{R} \subseteq \mathbf{S}$ is (ir)relevant, but does not discuss how to determine the partition of $\mathbf{S}$ into relevant and irrelevant nodes. Looking to facilitate finding such a partition we set out to study properties of (ir)relevant nodes. We first consider the relation to *computationally-irrelevant* nodes (Druzdzel and Suermondt, 1994) such as d-separated and *barren*[1] nodes and conclude that these can in fact be relevant according to Definition 1. MAP-independence therefore does not actually extend on conditional independence.

[1] Node V is barren if $V \in \mathbf{S}$, and V is a leaf node or a node with only barren descendants in the graph.

Proposition 3 *Let* $\mathbf{H}, \mathbf{E}, \mathbf{S}, \mathbf{R}$*, and* $\langle \mathbf{h}^*, \mathbf{e} \rangle$ *be as before. In the given context, let* $\mathbf{R}$ *be relevant to* $\mathbf{h}^*$*. Then* $\mathbf{R}' \subseteq \mathbf{R}$ can be *computationally-irrelevant in the given context.*

Proposition 3 poses a rather weak claim that can be substantiated through the example below; in our examples we exploit that

$$\arg\max_{\mathbf{h}' \in \Omega(\mathbf{H})} \Pr(\mathbf{h}'\,\mathbf{r} \mid \mathbf{e}) = \arg\max_{\mathbf{h}' \in \Omega(\mathbf{H})} \Pr(\mathbf{h}' \mid \mathbf{r}\,\mathbf{e}) \cdot \Pr(\mathbf{r} \mid \mathbf{e}) = \arg\max_{\mathbf{h}' \in \Omega(\mathbf{H})} \Pr(\mathbf{h}' \mid \mathbf{r}\,\mathbf{e})$$

Example 1 *Consider the Bayesian network in Figure 1a, with the following probabilities specified for nodes* A*,* B *and* C*:*

$$\begin{array}{lll} \Pr(a) = 0.55 & \Pr(c \mid a\,b) = 0.8 & \Pr(c \mid a\,\overline{b}) = 0.55 \\ \Pr(b) = 0.6 & \Pr(c \mid \overline{a}\,b) = 0.1 & \Pr(c \mid \overline{a}\,\overline{b}) = 0.5 \end{array}$$

Let $\mathbf{E} = \emptyset$ *and* $\mathbf{H} = \{A\}$*. Then, node* C *is a barren node and node* B *is d-separated from node* A*, so both* B *and* C *are computationally-irrelevant to* $\mathbf{H}$ *given* $\mathbf{E}$*. However, both* $\{C\}$ *and* $\{B, C\}$ *are relevant to* $\mathbf{h}^* = a$*:* $\Pr(\overline{a} \mid \overline{c}) = 0.67$ *and whereas* $\Pr(a \mid \overline{b}\,\overline{c}) = 0.52$*,* $\Pr(\overline{a} \mid b\,\overline{c}) = 0.79$*.* $\{B\}$ *is irrelevant to* $\mathbf{h}^*$ *but becomes relevant in combination with* $\overline{c}$ *!*

In the above example both set $\{C\}$ and its superset $\{B, C\}$ are relevant. In general, any superset of a relevant set is relevant.

Proposition 4 *Let* $\mathbf{H}, \mathbf{E}, \mathbf{S}$*, and* $\langle \mathbf{h}^*, \mathbf{e} \rangle$ *be as before, and let* $\mathbf{R}' \subseteq \mathbf{R} \subseteq \mathbf{S}$*. In the given context, if* $\mathbf{R}'$ *is relevant to* $\mathbf{h}^*$ *then* $\mathbf{R}$ *is relevant to* $\mathbf{h}^*$*.*

An equivalent proposition states that $\mathbf{h}^*$ is MAP-independent of $\mathbf{R}'$ if $\mathbf{h}^*$ is MAP-independent of $\mathbf{R}$ and is proven by Valero-Leal et al. (2022). Whereas every subset of an irrelevant set is irrelevant, not every union of irrelevant sets remains irrelevant.

Proposition 5 *Let* $\mathbf{H}, \mathbf{E}, \mathbf{S}$*, and* $\langle \mathbf{h}^*, \mathbf{e} \rangle$ *be as before. Let* $\mathbf{R}$*,* $\mathbf{R}' \subseteq \mathbf{S}$ *both be* ir*relevant to* $\mathbf{h}^*$*. Then* $\mathbf{R} \cup \mathbf{R}'$ can be *relevant to* $\mathbf{h}^*$*.*

We again substantiate this claim through an example.

Example 2 *Consider the Bayesian network in Figure 1a, with the following probabilities specified for nodes* A*,* B*,* C *and* D*:*

$$\begin{array}{llll} \Pr(a) = 0.3 & \Pr(c \mid a\,b) = 0.5 & \Pr(c \mid a\,\overline{b}) = 0.7 & \Pr(d \mid c) = 0.25 \\ \Pr(b) = 0.5 & \Pr(c \mid \overline{a}\,b) = 0.4 & \Pr(c \mid \overline{a}\,\overline{b}) = 0.8 & \Pr(d \mid \overline{c}) = 0.7 \end{array}$$

Let $\mathbf{E} = \{B\}$ *with* $\mathbf{e} = \overline{b}$*, and let* $\mathbf{H} = \{C\}$*. Then* $\mathbf{h}^* = c$*, and both* $\{A\}$ *and* $\{D\}$ *are irrelevant to* c*:* $\Pr(c \mid \overline{a}) = 0.8$*,* $\Pr(c \mid a) = 0.7$*,* $\Pr(c \mid \overline{d}) = 0.89$*, and* $\Pr(c \mid d) = 0.54$*. However,* $\Pr(c \mid a\,d) = 0.45$ *and therefore* $\{A, D\}$ *is relevant to* c*.*

From Propositions 4 and 5 we have that a partition of $\mathbf{S}$ into a set $\mathbf{S}^+$ of relevant nodes and a set $\mathbf{S}^-$ of irrelevant nodes, as suggested by Kwisthout (2021), is perhaps not what we should be looking for. It seems more appropriate to partition at the level of subsets, i.e. partitioning the powerset $\mathcal{P}(\mathbf{S})$ into a set $\mathcal{P}^+(\mathbf{S})$ of relevant subsets of $\mathbf{S}$, which we call *relevance sets*, and a set $\mathcal{P}^-(\mathbf{S})$ of irrelevant subsets of $\mathbf{S}$. Unfortunately, this observation makes the problem of finding relevance sets only harder.

4. Finding Relevant Nodes

To provide explanations that are as simple as possible, we aim for small sets of relevant nodes. Moreover, the computational costs of establishing the relevance of a given set $\mathbf{R}$ depends heavily on its size (Kwisthout, 2021). From the examples and propositions in the previous section, we have that a set of relevant nodes can contain irrelevant subsets that may or may not contribute to the relevance of the superset. To establish a minimal set of relevant nodes, or equivalently a maximal set of irrelevant nodes, we therefore cannot simply discard nodes that are individually irrelevant from a set of relevant nodes. In this section we explore several approaches to finding sets of (ir)relevant nodes. First we will consider finding nodes that are individually relevant, before discussing the more general case.

4.1 Singleton Relevance sets

To decide upon the relevance of *individual* nodes to the MAP-explanation, we first revisit the computationally-irrelevant nodes. We conclude that d-separated nodes cannot be relevant in isolation; as a result, they can be pruned from the network (Baker and Boult, 1991):

Proposition 6 *Let* $\mathbf{H}, \mathbf{E}, \mathbf{S}$*, and* $\langle \mathbf{h}^*, \mathbf{e} \rangle$ *be as before. In the given context, if node* $R \in \mathbf{S}$ *is d-separated from* $\mathbf{H}$ *given* $\mathbf{E}$ *then* $\{R\}$ *is irrelevant to* $\mathbf{h}^*$*.*

Proof This result is trivial: there are no active chains between R and $H_i \in \mathbf{H}$ given $\mathbf{E}$. ■

Individual barren nodes are also computationally-irrelevant, but can nonetheless be relevant to the MAP-explanation, as demonstrated in Example 1. We argue that relevant barren nodes can be useful for conveying information about the possible instability of a MAP-explanation. Consider for example a naive-Bayesian network classifier: every unobserved input node is a barren node. If such a node is relevant to the most-likely value of the class node, then it can be used to describe a context in which the classifier outcome may be different. We therefore suggest *not* to disregard barren nodes as possible relevant nodes.

A naive approach to establishing all relevant singletons would be to iterate over all $R \in \mathbf{R}$ and $r \in \Omega(R)$ and compute the MAP-explanation for evidence $r\,\mathbf{e}$. Instead, we propose to iterate over $\mathbf{h}' \in \Omega(\mathbf{H})$ and compute posterior distributions $\Pr(R \mid \mathbf{h}'\,\mathbf{e})$ using standard (join-tree) inference[2] and exploit the following property:

Proposition 7 *Let* $\mathbf{H}, \mathbf{E}, \mathbf{S}$*, and* $\langle \mathbf{h}^*, \mathbf{e} \rangle$ *be as before. Then node* $R \in \mathbf{S}$ *is relevant to* $\mathbf{h}^*$ *iff* $\exists \mathbf{h}' \in \Omega(\mathbf{H})$*,* $\mathbf{h}' \neq \mathbf{h}^*$ *such that*

$$\exists r \in \Omega(R) : \Pr(r \mid \mathbf{h}^*\,\mathbf{e}) = 0 \quad \textit{or} \quad \log \frac{\Pr(r \mid \mathbf{h}'\,\mathbf{e})}{\Pr(r \mid \mathbf{h}^*\,\mathbf{e})} + \log \frac{\Pr(\mathbf{h}'\,\mathbf{e})}{\Pr(\mathbf{h}^*\,\mathbf{e})} > 0$$

Proof R is relevant to $\mathbf{h}^*$ if $\exists r \in \Omega(R)$ such that $\Pr(\mathbf{h}'\,r \mid \mathbf{e}) > \Pr(\mathbf{h}^*\,r \mid \mathbf{e})$ for some value $\mathbf{h}' \in \Omega(\mathbf{H})$, $\mathbf{h}' \neq \mathbf{h}^*$. That is, either $\Pr(\mathbf{h}^*\,r \mid \mathbf{e}) = 0$ or

$$\frac{\Pr(\mathbf{h}'\,r \mid \mathbf{e})}{\Pr(\mathbf{h}^*\,r \mid \mathbf{e})} = \frac{\Pr(r \mid \mathbf{h}'\,\mathbf{e}) \cdot \Pr(\mathbf{h}' \mid \mathbf{e})}{\Pr(r \mid \mathbf{h}^*\,\mathbf{e}) \cdot \Pr(\mathbf{h}^* \mid \mathbf{e})} > 1$$

■

[2] One full network propagation serves to compute $\Pr(V \mid \mathbf{h}'\,\mathbf{e})$ for all $V \in \mathbf{V}$ (Jensen and Nielsen, 2007).

Algorithm 1: computing relevant singletons for context $\langle \mathbf{h}^*, \mathbf{e} \rangle$.

```
Input  : pruned BN $\mathcal{B}$, $\langle \mathbf{h}^*, \mathbf{e} \rangle$, $\mathbf{c_i} = \log[\Pr(\mathbf{h_i}\,\mathbf{e})/\Pr(\mathbf{h}^*\,\mathbf{e})]$ for all $\mathbf{h_i} \neq \mathbf{h}^*$
Output: Set $\mathcal{R}$ with relevant singletons
$\mathcal{R} \leftarrow \emptyset$; $\mathcal{S} \leftarrow \mathbf{S}$;
ComputePosteriors($\mathcal{B}, \mathbf{h}^*\,\mathbf{e}$);
forall $R \in \mathcal{S}$ do
    if $\exists r \in \Omega(R) : \Pr(r \mid \mathbf{h}^*\,\mathbf{e}) = 0$ then $\mathcal{R} \leftarrow \mathcal{R} \cup \{R\}$; $\mathcal{S} \leftarrow \mathcal{S} \setminus \{R\}$
end
forall $\mathbf{h_i} \neq \mathbf{h}^*$ do
    if $\mathcal{S} \neq \emptyset$ then ComputePosteriors($\mathcal{B}, \mathbf{h_i}\,\mathbf{e}$);
    forall $R \in \mathcal{S}$ do
        if $\exists r \in \Omega(R) : \log \frac{\Pr(r|\mathbf{h_i}\,\mathbf{e})}{\Pr(r|\mathbf{h}^*\,\mathbf{e})} + \mathbf{c_i} > 0$ then $\mathcal{R} \leftarrow \mathcal{R} \cup \{R\}$; $\mathcal{S} \leftarrow \mathcal{S} \setminus \{R\}$
    end
end
return $\mathcal{R}$
```

If we assume that the fractions $\Pr(\mathbf{h}'\,\mathbf{e})/\Pr(\mathbf{h}^*\,\mathbf{e})$ are available from the original MAP computations, we can exploit the above by propagating $\mathbf{h}'$ as additional evidence through the network and then labelling any node R that has a value r for which the property holds as relevant to $\mathbf{h}^*$. Any node labelled as relevant is indeed relevant and after all $\mathbf{h}' \neq \mathbf{h}^*$ have been propagated, we are guaranteed to have identified all relevant singletons. Pseudocode for this method is provided by Algorithm 1.

Example 3 (Experiment) *We performed the described procedure on the Insurance network shown in Figure 1b. We entered evidence for 5 nodes: Age = `senior`, SocioEcon = `uppermiddle`, DrivHist = `many`, HomeBase = `city`, MakeModel = `sportscar`. We consider three different sets of hypothesis nodes: Accident (H_1), RiskAversion (H_2) and their combination $\{H_1, H_2\}$. Given the evidence, their most likely values are $h_1^* =$ `none`, $h_2^* =$ `adventurous`, and this combination is also the most likely combination $\mathbf{h}^*_{\mathbf{1,2}}$ for $\{H_1, H_2\}$.*

In Table 1 we list, for each explanation context, the number of nodes d-separated from $\mathbf{H}$ given $\mathbf{E}$, the number of nuisance nodes[3], barren nodes and relevant nodes. For context $\langle h_1^*, \mathbf{e} \rangle$ all of the relevant nodes are barren nodes; the relevant node in context $\langle h_2^*, \mathbf{e} \rangle$ is a computationally relevant, non-nuisance node. The set of nodes that are individually relevant to $\mathbf{h}^*_{\mathbf{1,2}}$ happens to be the union of those relevant to h_1^* and to h_2^*.

For H_1 we find the same 6 relevant nodes for each of the values `mild`, `moderate` and `severe` $\in \Omega(H_1)$. In this case, we have therefore found all relevant nodes after only a single iteration of the algorithm. For H_2 we find the relevant node for 2 of the 3 values other than h_2^*. For $\{H_1, H_2\}$ we evaluate 15 configurations. In 6 of these configurations, we find the same 6 relevant nodes that we found for h_1^*; in 2 cases we find a subset of 5 of these 6 nodes, and in 5 cases a subset of 4 nodes. Finally, for 1 configuration we find 5 relevant

[3] $V \in \mathbf{S}$ is a nuisance node if it is computationally relevant, yet not on any active chain between $\mathbf{H}$ and $\mathbf{E}$; nuisance nodes should arguably not be included in explanations (Druzdzel and Suermondt, 1994).

context	# d-sep	# nuisance	# barren	# relevant
$\langle h_1^*, \mathbf{e} \rangle$	2	1	12	6
$\langle h_2^*, \mathbf{e} \rangle$	3	0	16	1
$\langle \mathbf{h}_{\mathbf{1,2}}^*, \mathbf{e} \rangle$	2	1	12	7

Table 1: For each explanation context from Example 3: the number of nodes d-separated from $\mathbf{H}$ given $\mathbf{E}$, the number of nuisance nodes, barren nodes and relevant singletons. Nodes that are both d-separated and barren are counted under # d-sep only.

nodes, 4 of which are a subset of those found for h_1^* and the 5th being the relevant node for h_2^*. Only for 1 configuration did we find no relevant nodes.

In our above experiment we observe that the nodes that are relevant to a most-likely combination of hypotheses is exactly the union of the nodes relevant to the individual most-likely hypotheses. If this property holds in general, then the number of propagations necessary to find all relevant singletons can be reduced from $|\Omega(\mathbf{H})|$ to $\sum_{H \in \mathbf{H}} |\Omega(H)|$. Unfortunately, this is not a general property, as is shown in the next example.

Example 4 *Consider the Bayesian network in Figure 1a, with the following probabilities specified for nodes A, B and C:*

$$\begin{array}{lll} \Pr(a) = 0.15 & \Pr(c \mid a\, b) = 0.4 & \Pr(c \mid a\, \bar{b}) = 0.7 \\ \Pr(b) = 0.65 & \Pr(c \mid \bar{a}\, b) = 0.5 & \Pr(c \mid \bar{a}\, \bar{b}) = 0.05 \end{array}$$

Let $\mathbf{E} = \emptyset$ and $\mathbf{H} = \{A, B\}$. The most likely value of A is $\bar{a}$, that of B is b; the most likely combination is also $\bar{a}\, b$. Node $\{C\}$ is irrelevant to both $\bar{a}$ and b, but relevant to the combination $\bar{a}\, b$: $\Pr(\bar{a}\, \bar{b} \mid \bar{c}) > \Pr(\bar{a}\, b \mid \bar{c})$.

4.2 Larger Relevance Sets

From Proposition 4 we have that any superset of the relevant singletons for $\mathbf{h}^*$ is also a relevance set for $\mathbf{h}^*$. For the purpose of explaining the robustness of a MAP-explanation, however, relevance sets should preferably not include nodes that do not actually contribute to the relevance. In this section we explore a number of options to establish larger relevance sets that are suitable for explanation purposes by discussing which nodes in $\mathbf{S}$ to consider for investigating relevance.

First, we reconsider the role of nodes $D \in \mathbf{S}$ that are d-separated from $\mathbf{H}$ given $\mathbf{E}$. If a node $V \in \mathbf{S}$ is added to a relevance set that serves to unblock a chain between a d-separated node D and a hypothesis node, then node D is no longer d-separated from the hypothesis and in fact can become relevant; we have observed this in Example 1. This suggests that we should no longer disregard initially d-separated nodes when searching for larger relevance sets. However, we could still argue that nodes that are conditionally independent when computing the MAP-explanation should be considered irrelevant and disregard them for that reason: explaining (in)stability of the MAP in terms of independent nodes may be very counter-intuitive to the user.

Suggestion 1 *Prior to establishing relevance sets of any size for context $\langle \mathbf{h}^*, \mathbf{e} \rangle$, prune* $\mathbf{S}$ *by removing all nodes d-separated from* $\mathbf{H}$ *by* $\mathbf{E}$.

Even if all d-separated nodes are pruned as suggested, d-separation can still be a useful tool for reducing the size of established relevance sets. Consider a relevance set $\mathbf{R} \subseteq \mathbf{S}$ and let $R \in \mathbf{R}$ be relevant to $\mathbf{h}^*$. Then any node $U \in \mathbf{R}$, $U \neq R$, that is d-separated from $\mathbf{H}$ given $\{R\} \cup \mathbf{E}$ can be removed from $\mathbf{R}$ regardless of whether or not U is relevant to $\mathbf{h}^*$. The relevance of the singleton $\{U\}$ in this case was due to the relevance of R and in their combination U no longer contributes to changing the distribution over the hypotheses; the node is therefore superfluous in $\mathbf{R}$.

Suggestion 2 *Let* $\mathbf{R}$ *be a relevance set for context* $\langle \mathbf{h}^*, \mathbf{e} \rangle$*. If* $\mathbf{D} \subset \mathbf{R}$ *is d-separated from* $\mathbf{H}$ *by* $\mathbf{E} \cup \mathbf{R} \setminus \mathbf{D}$ *then* $\mathbf{R} \setminus \mathbf{D}$ *is more suitable for the purpose of explanation than* $\mathbf{R}$*.*

Recall from Proposition 5 that a combination of irrelevant nodes can become relevant. To prevent having superfluous nodes in a relevance set, which easily happens when considering supersets of relevant sets, we could restrict attention to combinations of irrelevant sets only. An option is to start by evaluating every pair of irrelevant nodes and expanding the sets as long as the combination remains irrelevant. If we have n irrelevant singletons, then we already have $n \cdot (n-1)/2$ pairs to evaluate. In the contexts we experimented with for the Insurance BN in Example 3, for example, we found at most 7 relevant singletons, leaving 10 irrelevant non-d-separated supplementary nodes and therefore 45 pairs to explore. Determining the irrelevance of these pairs is moreover more complex than the approach from Algorithm 1 since we need to compute posteriors over two nodes. Ultimately, we end up solving the MAP-independence problem a large number of times.

To reduce the number of irrelevant sets to evaluate, again d-separation may be employed. The following proposition, but phrased in terms of MAP-independence and conditional independence, is proven by Valero-Leal et al. (2022).

Proposition 8 *Let* $\mathbf{H}, \mathbf{E}, \mathbf{S}$*, and* $\langle \mathbf{h}^*, \mathbf{e} \rangle$ *be as before, and let* $\mathbf{U}, \mathbf{W} \subset \mathbf{S}$*. In the given context, if* $\mathbf{U}$ *is irrelevant to* $\mathbf{h}^*$ *and* $\mathbf{W}$ *is d-separated from* $\mathbf{H}$ *given* $\mathbf{U} \cup \mathbf{E}$ *then* $\mathbf{W}$ *is irrelevant to* $\mathbf{h}^*$*.*

A set that d-separates a node from every other node in the graph is the *Markov boundary* MB() of the node, consisting of the node's parents, children, and co-parents[4] (Pearl, 1988). If the Markov boundary of a hypothesis node H were to be irrelevant to the most likely value h^* of the node, given the evidence, then the whole of $\mathbf{S}$ is irrelevant to h^*. Establishing the (ir)relevance of the Markov boundary of H can be done by inspecting the conditional probability distributions specified locally for H and its children.

Proposition 9 *Let* $\mathbf{H} = \{H\}$ *and* $\langle h^*, \mathbf{e} \rangle$ *be as before and let* $\sigma(H)$ *denote the set of all children of* H*. Then* $\mathbf{M} = \text{MB}(H)$ *is irrelevant to* h^* *iff* $\forall \mathbf{m} \in \Omega(\mathbf{M})$ *and* $\forall h' \neq h^* \in \Omega(H)$*:*

$$\Pr(h^* \, \mathbf{m} \mid \mathbf{e}) > 0 \quad \textit{and} \quad \sum_{C \in \sigma(H)} \log \frac{\Pr(c \mid h' \, \pi_{C \setminus H})}{\Pr(c \mid h^* \, \pi_{C \setminus H})} + \log \frac{\Pr(h' \mid \pi_H)}{\Pr(h^* \mid \pi_H)} \leq 0$$

where configurations $c \in \Omega(C)$*,* $\pi_{C \setminus H} \in \Omega(\pi(C) \setminus \{H\})$ *and* $\pi_H \in \Omega(\pi(H))$ *are consistent with* $\mathbf{m}$*. If* $\text{MB}(H) \cap \mathbf{E} \neq \emptyset$*, then we need to consider only those* $\mathbf{m}$ *consistent with* $\mathbf{e}$*.*

[4] Note that these sets are not necessarily disjunct.

Proof By definition, $\textsc{mb}(H)$ is irrelevant to h^* if $\forall \mathbf{m} \in \Omega(\mathbf{M})$ and $\forall h' \neq h^* \in \Omega(H)$

$$\Pr(h' \, \mathbf{m} \mid \mathbf{e}) \leq \Pr(h^* \, \mathbf{m} \mid \mathbf{e})$$

Assume each $\mathbf{m}$ is consistent with $\mathbf{e}$ and let $\mathbf{m} = \mathbf{p}\,\mathbf{c}\,\mathbf{s}$, where $\mathbf{p}$ is the configuration for the parents of H, $\mathbf{c}$ that for the children and $\mathbf{s}$ that of the spouses (co-parents) of H. Then

$$\frac{\Pr(h' \, \mathbf{m} \mid \mathbf{e})}{\Pr(h^* \, \mathbf{m} \mid \mathbf{e})} = \frac{\Pr(h' \mid \mathbf{m}\,\mathbf{e})}{\Pr(h^* \mid \mathbf{m}\,\mathbf{e})} = \frac{\Pr(h' \mid \mathbf{m})}{\Pr(h^* \mid \mathbf{m})} = \frac{\Pr(\mathbf{c} \mid h' \, \mathbf{p}\,\mathbf{s}) \cdot \Pr(h' \mid \mathbf{p}\,\mathbf{s})}{\Pr(\mathbf{c} \mid h^* \, \mathbf{p}\,\mathbf{s}) \cdot \Pr(h^* \mid \mathbf{p}\,\mathbf{s})} \leq 1$$

The result follows by using the fact that all children are mutually independent given H, its parents and co-parents. ■

We can employ Proposition 9 also in case $\mathbf{H}$ contains more than one node.

Proposition 10 *Let $\langle \mathbf{h}^*, \mathbf{e} \rangle$ be as before. Let $\mathbf{H} = \{H_1, \ldots, H_n\}$, $n > 1$, and consider the additional explanation contexts $\langle h_i^*, \mathbf{e} \rangle$, $h_i^* \in \Omega(H_i)$, for each $H_i \in \mathbf{H}$. If for each H_i, $\textsc{mb}(H_i)$ is irrelevant to h_i^* then $\mathbf{h}^* = h_1^* \ldots h_n^*$ and $\cup_{i=1}^{n} \textsc{mb}(H_i)$ is irrelevant to $\mathbf{h}^*$.*

Proof Follows directly from independence properties given the Markov boundaries. ■

Investigating the (ir)relevance of Markov boundaries, although relatively efficient since all computations can be done locally, has several drawbacks. First of all, larger Markov boundaries, are less likely to be irrelevant, as is clear from the next example.

Example 5 *Reconsider the Insurance BN and the same evidence as in Example 3. We examined the relevance of the Markov boundary for each of the 22 other nodes in the graph, taking the evidence into account. One node, Good Student, is completely determined by the evidence and can therefore not change value. For 19 nodes the Markov blanket is relevant to their most likely value; for 13 of these the conditional probability distribution specified for the node itself already suffices to draw that conclusion. Only for two nodes, IliCost and OtherCar, is their Markov boundary irrelevant to their most likely value. Note that both these nodes have only a single node in their Markov boundary.*

If not the entire Markov boundary of H is irrelevant, we could determine the irrelevant subset $\mathbf{R}'$ and subsequently add the Markov boundaries of the nodes $\textsc{mb}(H) \setminus \mathbf{R}'$ to $\mathbf{R}'$ and evaluate their irrelevance. Again this may quickly result in very large sets to evaluate while finding only few of them to be irrelevant. Moreover, another drawback of this approach is that as soon as we go outside the Markov boundaries, Proposition 10 no longer applies and we have already seen that we cannot in general aggregate the results for multiple hypothesis nodes to draw conclusions about (ir)relevance to a most likely combination of hypotheses. To somewhat alleviate the drawbacks of the two discussed approaches to establishing larger relevance sets, we suggest to employ a combination of the two, starting from irrelevant singletons in the Markov boundaries of the hypothesis nodes.

Suggestion 3 *Let $\mathcal{R}$ be the set of relevant singletons for context $\langle \mathbf{h}^*, \mathbf{e} \rangle$ as returned by Algorithm 1. Initialise $\mathcal{I} = \mathbf{V} \setminus (\mathbf{H} \cup \mathbf{E} \cup \mathcal{R})$ to the set of irrelevant singletons. Check each pair of nodes from $\mathcal{I}$ for their relevance, where we start with nodes from the Markov boundaries of the hypothesis nodes. If the pair is relevant to $\mathbf{h}^*$ then we add it to $\mathcal{R}$ and stop investigating its supersets, otherwise we increase the subset of $\mathcal{I}$ by one and investigate its relevance. We can stop this process any time we have found* enough *relevance sets of* sufficient *size for explanation purposes (see the discussion in Section 5).*

Note that, for reasons of efficiency, the suggested procedure does not investigate supersets of sets that are already established as relevant. As long as explanations only focus on the nodes in the relevance sets this is fine, since smaller sets provide for more simple explanations. If we want to include the specific observations for these nodes that render them relevant, however, then we may want to consider supersets: it could be the case that node R_1 is irrelevant to $\mathbf{h}^*$, node R_2 is relevant because $\overline{r}_2 \in \Omega(R_2)$ results in a MAP-explanation $\mathbf{h}' \neq \mathbf{h}^*$, and that the combination is relevant because $r_1\, r_2 \in \Omega(\{R_1, R_2\})$ together result in yet another MAP explanation.

Finally, since relevant nodes need to be observed in order to actually affect the MAP-explanation, we could choose to evaluate only observable nodes for their relevance.

Suggestion 4 *Let* $\mathbf{H}, \mathbf{E}, \mathbf{S}$*, and* $\langle \mathbf{h}^*, \mathbf{e} \rangle$ *be as before. Let* $\mathbf{O} \subseteq \mathbf{S}$ *be the set of observable nodes outside* $\mathbf{E}$*. Then only evaluate nodes in* $\mathbf{R} \subseteq \mathbf{O}$ *for their (ir)relevance.*

5. Using Relevance in Explanations

The sets of relevant nodes found can be used to explain the potential (in)stability of a MAP-explanation. We can explain the user that the MAP-explanation is expected to remain the same as long as relevant nodes are not observed. Based upon user preferences we can restrict the explanation to relevant singletons or also return larger sets. In addition, if we define a measure of *degree of relevance*, the relevance sets can be ordered from most relevant to least relevant to the stability of the MAP-explanation. A degree of relevance can be defined in various ways. We can define an *expected relevance* to capture how likely it is that the MAP-explanation will change due to future observations for $\mathbf{R}$:

$$\mathit{ExpRel}(\mathbf{R}, \mathbf{h}^*, \mathbf{e}) = \sum_{\mathbf{r} \in \text{REL}(\mathbf{R}, \langle \mathbf{h}^*, \mathbf{e} \rangle)} \Pr(\mathbf{r} \mid \mathbf{e}),$$

where $\text{REL}(\mathbf{R}, \langle \mathbf{h}^*, \mathbf{e} \rangle) = \{\mathbf{r} \in \Omega(\mathbf{R}) \mid \arg\max_{\mathbf{H}} \Pr(\mathbf{H}\, \mathbf{r} \mid \mathbf{e}) = \mathbf{h}' \text{ for any } \mathbf{h}' \neq \mathbf{h}^*\}$. *ExpRel* is similar in concept to MAP-*independence strength* (Valero-Leal et al., 2022) and the *same-decision probability* (Choi et al., 2012). Another option is to consider the average probability mass assigned to an alternative MAP explanation in case of future observations for $\mathbf{R}$:

$$\mathit{AvgRel}(\mathbf{R}, \mathbf{h}^*, \mathbf{e}) = \frac{1}{|\text{REL}(\mathbf{R}, \langle \mathbf{h}^*, \mathbf{e} \rangle)|} \sum_{\mathbf{r} \in \text{REL}(\mathbf{R}, \langle \mathbf{h}^*, \mathbf{e} \rangle)} \Pr(\mathbf{h}' \mid \mathbf{r}\, \mathbf{e})$$

In addition to indicating which nodes are relevant, and possibly their degree of relevance, we can keep track of the specific values $\mathbf{r}$ of the nodes that are relevant to the MAP-explanation $\mathbf{h}^*$ and the value $\mathbf{h}'$ into which the most-likely hypothesis changes upon their observation. This allows us to provide contrastive and counterfactual robustness explanations, where

- any $\mathbf{r}$ found provides for a *counterfactual* explanation for the instability of $\mathbf{h}^*$;
- any $\mathbf{r}$ found that results in $\mathbf{h}'$ as most likely value provides for a robustness explanation that *contrasts* outputs $\mathbf{h}^*$ and $\mathbf{h}'$.

We would like to note that contrastive and counterfactual explanations usually focus on changes in the observed value of the evidence nodes, rather than on effects of possible future observations (see e.g. Koopman and Renooij (2021)). Moreover, the term 'counterfactual' here does not have a causal connotation as it often does have (Halpern and Pearl, 2005).

The above discussion shows there are ample possibilities to use relevance sets and associated configurations to explain the (in)stability of a MAP-explanation to a user. Moreover, it shows that we can to some extent meet the explanation criteria listed by Miller (2019). Relevance computations provide information for generating contrastive explanations, indicating the additional evidence that is required to obtain an expected MAP-explanation $\mathbf{h}'$ rather than the current MAP $\mathbf{h}^*$. Rather than returning all relevance sets, the explanation can focus on small relevance sets, or even just the singletons. Using a measure of degree of relevance we can moreover use only the most relevant ones in the robustness explanation, or at least prioritize the explanations based upon degree of relevance. By interacting with the user, taking into account their preference with respect to size and number of relevance sets to include in the robustness explanation, the explanation becomes social. Finally, although it is difficult to refrain from referring to probabilities or statistical relationships in the context of Bayesian networks, explaining the (in)stability of MAP-explanations can be done in terms of changes in the MAP rather than changes in probabilities.

6. Conclusions and Further Research

In this paper we have studied properties of MAP-independence and discussed approaches to finding relevance sets and using these in explaining the robustness of MAP-explanations. We have seen that combinations of irrelevant sets can become relevant and that any superset of a relevant set is relevant, yet may contain superfluous nodes that do not really contribute to the relevance. These properties make it difficult to simply partition the set of supplementary nodes $\mathbf{S}$ into a relevant set and an irrelevant set. Rather, we have different relevant subsets and various irrelevant subsets, suggesting that a partition of the powerset of $\mathbf{S}$ is required. This shows the complexity of the problem of finding relevance sets.

We have proposed an algorithm for finding relevant singletons and have suggested some simple approaches to finding non-singleton relevance sets that have several drawbacks. In future we would like to investigate and detail more efficient approaches to this end. Moreover, we will study and experiment with the proposed measures of degree of relevance in more detail. Finally, we would like to evaluate the suggested explanations with actual users.

Acknowledgments

All Bayesian network computations were done using GeNIe (www.bayesfusion.com/genie/); the Insurance BN was taken from the bnlearn repository (www.bnlearn.com/bnrepository/). This research was supported by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, `https://hybrid-intelligence-centre.nl`.

References

E. Albini, A. Rago, P. Baroni, and F. Toni. Influence-driven explanations for Bayesian network classifiers. In *Proceedings of the Eighteenth Pacific Rim International Conference on Artificial Intelligence*, LNAI 13031, pages 88–100. Springer Nature, Switzerland, 2021.

M. Baker and T. E. Boult. Pruning Bayesian networks for efficient computation. In *Uncertainty in Artificial Intelligence 6*, pages 225–232. Elsevier Science, Amsterdam, 1991.

J. Binder, D. Koller, S. Russell, and K. Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244, 1997.

A. Choi, Y. Xue, and A. Darwiche. Same-decision probability: A confidence measure for threshold-based decisions. *International Journal of Approximate Reasoning*, pages 1415–1428, 2012.

M. J. Druzdzel and H. Suermondt. Relevance in probabilistic models: backyards in a small world. In *Working Notes of the AAAI 1994 Fall Symposium Series: Relevance*, pages 60–63, 1994.

J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach. Part II: Explanations. *The British Journal for the Philosophy of Science*, 56(4):889–911, 2005.

F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs.* Springer Science & Business Media, 2nd edition, 2007.

T. Koopman and S. Renooij. Persuasive contrastive explanations for Bayesian networks. In *Proceedings of the Sixteenth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, LNAI 12897, pages 229–242. Springer, 2021.

J. Kwisthout. Explainable AI using MAP-independence. In *Proceedings of the Sixteenth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, LNAI 12897, pages 243–254. Springer, 2021.

C. Lacave and F. J. Díez. A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127, 2002.

T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems. Networks of Plausible Inference.* Morgan Kaufmann, 1988.

J. Pearl and A. Paz. GRAPHOIDS: a graph-based logic for reasoning about relevance relations. Technical Report R-53-L, UCLA Computer Science Department, 1987.

E. Valero-Leal, P. Larrañaga, and C. Bielza. Extending MAP-independence for Bayesian network explainability. In *Workshop on Heterodox Methods for Interpretable and Efficient Artificial Intelligence*, June 2022. URL `https://hmieai2022.cs.umu.se/`.