International Journal of Approximate Reasoning 160 (2023) 108965

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

# Efficient search for relevance explanations using MAP-independence in Bayesian networks

Enrique Valero-Leal [a], Concha Bielza [a], Pedro Larrañaga [a], Silja Renooij [b,*]

[a] Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Spain
[b] Department of Information and Computing Sciences, Utrecht University, the Netherlands

ARTICLE INFO

*Article history:*
Received 14 February 2023
Received in revised form 12 May 2023
Accepted 12 June 2023
Available online 15 June 2023

*Keywords:*
Bayesian network
Relevance
Explainability
Map-independence
Defeasible reasoning
Robustness

ABSTRACT

MAP-independence is a novel concept concerned with explaining the (ir)relevance of intermediate nodes for maximum a posteriori (MAP) computations in Bayesian networks. Building upon properties of MAP-independence, we introduce and experiment with methods for finding sets of relevant nodes using both an exhaustive and a heuristic approach. Our experiments show that these properties significantly speed up run time for both approaches. In addition, we link MAP-independence to defeasible reasoning, a type of reasoning that analyses how new evidence may invalidate an already established conclusion. Ways to present users with an explanation using MAP-independence are also suggested.



## 1. Introduction

In recent years, interest in explainable artificial intelligence (XAI) has grown, as it might be a solution to tackle the performance, ethical, and legal problems raised by deep learning models, such as neural networks. Among the possible approaches being researched to solve such issues is the use of transparent models rather than black-boxes, since they can be a better option in terms of explainability [29]. In this paper, our main focus is on explanations under uncertainty. In order to capture uncertainty in a model that is considered to be transparent, we use Bayesian networks, which compactly represent a joint probability distribution through a directed acyclic graph and associated conditional probability tables [14].

There are various ways to tackle the problem of explaining Bayesian networks, classified into different categories according to what is explained (see Fig. 1 for a taxonomy on the matter). Our previous work focused more on how to present a network or the underlying reasoning to the user, either using argumentation as an alternative [34] or by simplifying massive temporal networks [37]. In this paper, however, we focus on domain-independent and general methods for explaining the evidence, specifically for the maximum a posteriori (MAP) explanation [20]. A MAP-explanation is the outcome of computing the 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 marginalised 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

* Corresponding author.
*E-mail addresses:* enrique.valero@upm.es (E. Valero-Leal), mcbielza@fi.upm.es (C. Bielza), pedro.larranaga@fi.upm.es (P. Larrañaga), s.renooij@uu.nl (S. Renooij).

https://doi.org/10.1016/j.ijar.2023.108965

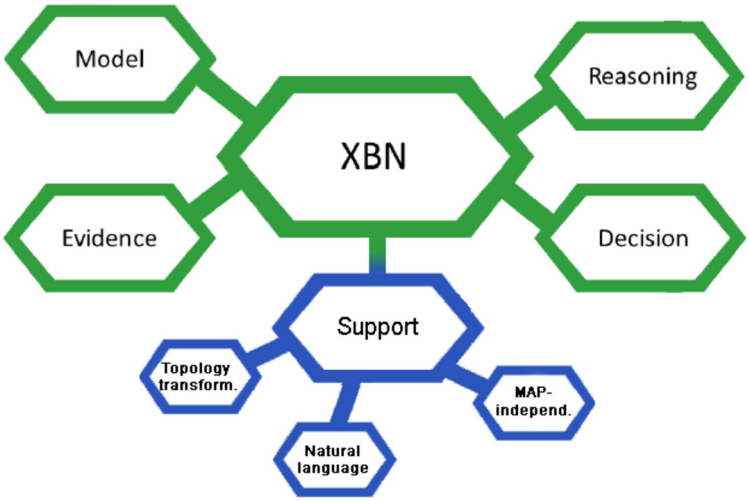


**Fig. 1.** Taxonomy of explanations in Bayesian networks (adaptation of figure from [9]). The Support category, in blue, captures our contributions presented in previous works. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

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 [17,18]. In Bayesian network literature, the word *(ir)relevance* has been used for various purposes, to denote different properties of different sets of nodes. In this paper, we will compare a few of these concepts; for a more exhaustive discussion, see [17]. 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 XAI.

Many existing evidence explanation methods for Bayesian networks focus on directly relating input (evidence) and output in the context of Bayesian classifiers, (see e.g. [16] for a brief overview). Recognising 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 [1]. In fact, Bayesian network classifiers [4] typically return a MAP-explanation as output, and we see MAP-independence as an alternative for including intermediate nodes in the explanation in networks where no structural assumptions are made. The concept of MAP-independence, while introduced for networks with discrete nodes only, has also been generalised to networks with continuous variables [36].

In addition to using node relevance in explanations for Bayesian networks and as a tool for explaining the stability of a result, we argue that we can link the concept of MAP-independence to defeasible reasoning, a branch of argumentation that provides for reasoning under uncertainty by allowing new premises to alter a previously established conclusion. Since good explanations are preferably simple, social, contrastive and do not refer to probabilities [24], we believe that linking our ideas to concepts of logic and argumentation can give our work a wider logical grounding, similar to how contrastive and counterfactual reasoning is used for explainability purposes [32].

Determining whether a MAP-explanation is MAP-independent of a *given* subset $\mathbf{R}$ of intermediate nodes is a co-NP$^{\text{PP}}$-complete problem [18]. In this paper, our main focus is on efficiently determining *which* subsets $\mathbf{R}$ are of interest to investigate in terms of MAP-independence. In doing so, we combine and extend the theoretical results presented in our earlier conference papers [28,36]. More specifically:

- We further elaborate on the properties of MAP-independence, which enable us to propose more optimal approaches to finding relevant subsets and extend on the proposal for singleton sets of [28];
- we experimentally test the algorithms presented in [28] and introduce two new ones: an exhaustive and a heuristic approach for finding larger relevant sets of intermediate nodes;
- the proposed approaches are studied and contrasted experimentally, testing them against different benchmark networks.

The paper is organised as follows. We present some relevant preliminaries, terminology and notation in Section 2. Section 3 defines (ir)relevance and its properties. Approaches to finding relevant sets are proposed in Section 4 and ideas to use them as explanations for the user are discussed in Section 5. In Section 6, we experimentally test our proposals, upon which the paper is concluded in Section 7.

## 2. Preliminaries

In this section, we provide some relevant background on Bayesian networks, notions of (ir)relevance, defeasible reasoning, and introduce our notation.

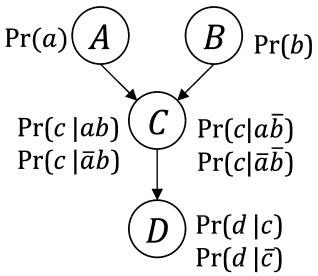


**Fig. 2.** A simple Bayesian network.

### 2.1. Bayesian networks

We consider a Bayesian network $\mathcal{B} = (G, \Pr)$ representing a joint probability distribution $\Pr(\mathbf{V})$ over a set of discrete random variables $\mathbf{V}$ [25]. We use capital letters to denote variables, typically bold-faced in the case of sets. We use $\Omega(V)$ to represent the domain of the variable $V \in \mathbf{V}$, writing $v$ as the 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 $\Omega(\mathbf{V}) = \times_{V \in \mathbf{V}} \Omega(V)$. $|\mathbf{V}|$ denotes the size of set $\mathbf{V}$.

Each variable is represented by a node in the directed acyclic graph (DAG) $G$, which captures the conditional (in)dependence relations in $\mathbf{V}$ through the notion of *d-separation* [25]. Two nodes $X$ and $Y$ are d-separated by a set of nodes $\mathbf{Z}$ iff every chain in $G$ between $X$ and $Y$ is *blocked* (inactive), that is, the chain contains a node $Z \in \mathbf{Z}$ with at most one incoming arc on the chain, or a node with two incoming arcs that is not in $\mathbf{Z}$ and has no descendants in $\mathbf{Z}$. If $X$ and $Y$ are d-separated by $\mathbf{Z}$ then the associated variables $X$ and $Y$ are conditionally independent given $\mathbf{Z}$. With each variable, or node, $V$ a set of local probability distributions $\Pr(V \mid \boldsymbol{\pi}(V))$ is associated, one for each configuration of parents $\boldsymbol{\pi}(V)$ of $V$ in the DAG, which together define the joint probability distribution: $\Pr(\mathbf{V}) = \prod_{V \in \mathbf{V}} \Pr(V \mid \boldsymbol{\pi}(V))$. Fig. 2 shows an example Bayesian network with four binary-valued nodes and their required probability parameters (complements omitted).

In this work, we focus in explaining the (ir)relevance of intermediate nodes with respect to an outcome of interest. To this end, we assume that $\mathbf{V}$ is partitioned into three disjoint subsets: 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. Therefore, set $\mathbf{E}$ contains only the currently observed nodes and not necessarily all observable ones.

Inference in a Bayesian network amounts to computing probabilities of the form $\Pr(\mathbf{h} \mid \mathbf{e})$ for some $\mathbf{h} \in \Omega(\mathbf{H})$ and $\mathbf{e} \in \Omega(\mathbf{E})$ with $\Pr(\mathbf{e}) > 0$. Note that since

$$\Pr(\mathbf{h} \mid \mathbf{e}) = \frac{\Pr(\mathbf{h}\,\mathbf{e})}{\Pr(\mathbf{e})} \propto \sum_{\mathbf{s} \in \Omega(\mathbf{S})} \Pr(\mathbf{h}\,\mathbf{s}\,\mathbf{e}),$$

the supplementary nodes are the nodes that need to be summed out in the computation. Standard algorithms for (exact) inference are based on join-tree propagation, where one full network propagation serves to compute $\Pr(V \mid \mathbf{e})$, for any $V \in \mathbf{V}$ [14]. Standard inference (its decision variant[1]) is a PP-complete problem [19].

Our outcome of interest is a yet more complex query: a MAP-explanation $\text{MAP}(\mathbf{H}, \mathbf{e})$, i.e., a *most likely* hypothesis $\mathbf{h}^* \in \Omega(\mathbf{H})$ to explain the evidence $\mathbf{e} \in \Omega(\mathbf{E})$:

$$\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e}) = \arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h} \mid \mathbf{e}) = \arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h}\,\mathbf{e}).$$

We note that if the posterior distribution $\Pr(\mathbf{H} \mid \mathbf{e})$ is multi-modal, then $\mathbf{h}^*$ need not be unique since multiple most likely values can exist. Depending on the context of use, we can in such a case choose to have the arg max function return all of the MAP-explanations rather than a single one.

The decision variant of the MAP[2] is an $\text{NP}^{\text{PP}}$-complete problem [19]. We will refer to a tuple $\langle \mathbf{h}^*, \mathbf{e} \rangle$ as the *explanation context*. Since we do not assume the Bayesian network to be a causal model, the MAP-explanation is *not* necessarily a *causal* explanation for the observed evidence [13].

[1] Is $\Pr(\mathbf{h} \mid \mathbf{e}) > t$ for a given threshold $t$?

[2] Does an $\mathbf{h}$ exist with $\Pr(\mathbf{h} \mid \mathbf{e}) > t$ for a given threshold $t$?

### 2.2. Notions of irrelevance

As mentioned in Section 1, the concept of relevance has various interpretations in the context of Bayesian networks. In this section, we review several sets of nodes that have been identified as irrelevant for different purposes and that are interesting for the remainder of this paper.

From the DAG over $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$, several types of node can be distinguished, where the first two are computationally irrelevant (its parameters are not needed to compute the query and will not affect the output) and can be pruned from the network prior to performing inference for a query over $\mathbf{H}$ and $\mathbf{E}$ [3]:

1. Nodes $\mathbf{D}$ that are d-separated from $\mathbf{H}$ given $\mathbf{E}$;
2. Nodes $\mathbf{B}$ that are *barren* with respect to $\mathbf{H}$ and $\mathbf{E}$, where a node $B$ is called barren if $B \in \mathbf{S}$, and $B$ is a leaf node or a node with only barren descendants in the DAG [31];
3. Nodes $\mathbf{N}$ that are *nuisance* nodes with respect to $\mathbf{H}$ and $\mathbf{E}$, where a node $N$ is called a nuisance node if $N \in \mathbf{S}$ is computationally relevant, yet not on any active chain between $\mathbf{H}$ and $\mathbf{E}$ [33].

Pearl & Paz [26] define irrelevance in terms of d-separation, and nuisance nodes are considered irrelevant for explaining the reasoning chains between evidence and hypothesis variables [10]. Kwisthout [17] argues that "...there is a sense in which a variable has an explanatory role [...] that goes beyond conditional (in)dependence". For the purpose of explaining the (ir)relevance of supplementary nodes for computing a MAP-explanation, Kwisthout [17,18] introduces the concept of MAP-independence.

**Definition 1** *(Kwisthout [17]).* Let $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ and $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ be as before. Then $\mathbf{h}^*$ is MAP-*independent* of $\mathbf{R} \subseteq \mathbf{S}$ given evidence $\mathbf{e}$, denoted MAP-*ind*$(\mathbf{R}, \mathbf{h}^*, \mathbf{e})$, if for all $\mathbf{r} \in \Omega(\mathbf{R})$, $\arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h}\,\mathbf{r} \mid \mathbf{e}) = \mathbf{h}^*$.

We note that MAP-independence is defined with respect to a given MAP-explanation $\mathbf{h}^*$. If $\Pr(\mathbf{H} \mid \mathbf{e})$ is multi-modal and multiple most likely values $\mathbf{h}^*$ exist, then it may be interesting to consider whether or not *all* of the most likely values of $\mathbf{H}$ are MAP-independent of $\mathbf{R}$. Likewise, even if $\mathbf{h}^*$ is unique, there could be value assignments for $\mathbf{H}$ in addition to $\mathbf{h}^*$ that maximise the posterior $\Pr(\mathbf{H}\,\mathbf{r} \mid \mathbf{e})$ for a given $\mathbf{r}$, in which case we may want to render $\mathbf{R}$ relevant. To cover these (rare) cases we can choose to have the arg max function return the set of all most likely values and subsequently decide MAP-independence based on equality of such sets. Without loss of generality, however, we assume in this paper that no two posterior probabilities are exactly equivalent and therefore most likely values are unique.

If $\mathbf{h}^*$ is *not* MAP-independent of a set $\mathbf{R}$, we will call $\mathbf{h}^*$ MAP-*de*pendent of $\mathbf{R}$. Any set to which $\mathbf{h}^*$ is MAP-independent is called a MAP-independent set; likewise for MAP-dependence.

An alternative definition of MAP-independence [18] is stated in terms of the equality

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

and does not explicitly refer to $\mathbf{h}^*$. Note, however, that if the above equality holds for all value assignments $\mathbf{r}$ and $\mathbf{r}'$, i.e., regardless of the value assignment to $\mathbf{R}$ we find the same most likely value $\mathbf{h}$, then we find the same most likely value after summing out $\mathbf{R}$, i.e., $\mathbf{h} = \mathbf{h}^*$. Moreover, if the above equality does not hold for all value assignments to $\mathbf{R}$, then there is at least one value assignment that results in a most likely value different from $\mathbf{h}^*$. In this paper we adopt the version that explicitly uses $\mathbf{h}^*$ and similarly rephrase the definition of *weak* MAP-independence.

**Definition 2** *(Kwisthout [18]).* Let $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ and $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ be as before. Then $\mathbf{h}^*$ is *weakly* MAP-*independent* of $\mathbf{R} \subseteq \mathbf{S}$ if for all $R \in \mathbf{R}$ and all $r \in \Omega(R)$, $\arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h}\,r \mid \mathbf{e}) = \mathbf{h}^*$.

Another concept related to MAP-independence introduced by Kwisthout [18] highlights the importance of finding a maximum set $\mathbf{R}$ such that the explanation $\mathbf{h}^*$ is MAP-independent of $\mathbf{R}$. A maximum set in this case means that no larger subset of supplementary nodes adheres to the definition of MAP-independence. This concept is referred to as *maximum* MAP-independence and can be defined as follows:

**Definition 3** *(Kwisthout [18]).* Let $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ and $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ be as before and let $\mathbf{h}^*$ be MAP-independent of $\mathbf{R} \subseteq \mathbf{S}$. If there exists no $\mathbf{S}' \subseteq \mathbf{S}$ with $|\mathbf{S}'| > |\mathbf{R}|$ such that $\mathbf{h}^*$ is MAP-independent of $\mathbf{S}'$, then $\mathbf{h}^*$ is *maximum* MAP-independent of $\mathbf{R}$.

We note that the problem of (maximum) MAP-independence somewhat resembles the problem of feature subset selection (FSS) [30]. FSS can be defined as a task that aims towards keeping the relevant variables with respect to a target, while discarding those features that are irrelevant or redundant [22]. In FSS we can distinguish between filter and wrapper techniques. Filter techniques take into account only the intrinsic properties of the data, usually by means of computing a score between every variable (or a subset of variables) and the target or a class variable $C$. It is a fast and model-agnostic approach that often comes at the cost of performance, as these techniques tend to disregard the actual functioning of the

model. On the other hand, wrapper techniques repeatedly train and evaluate a classifier to assess the quality of subsets of variables, resulting in better performance but a higher computational burden.

### 2.3. Defeasible reasoning

In the introduction we suggested that MAP-independence can be linked to defeasible reasoning. To further detail the relation between the two concepts in Section 5 we provide some relevant background here.

Defeasible reasoning is a type of reasoning that studies the case in which, given premise $A$ and a rule that leads from $A$ to $C$, the proof of $C$ might not be complete, as there may be an argument $B$ that invalidates $C$ [15]. We then say that $B$ is a defeater for $C$. An example can be found in a scenario of predictive modelling in supervised classification where there is uncertainty over the predictor variables $\mathbf{X}$, divided into observable variables $\mathbf{X}_o$ and unobservable variables $\mathbf{X}_u$. If the prediction for the class (vector) $\mathbf{C}$ given a configuration $\mathbf{x}_o$ for the observable variables is $\mathbf{c}$, a subset $\mathbf{R} \subseteq \mathbf{X}_u$ would be a defeater if adding any configuration $\mathbf{r}$ to $\mathbf{x}_o$ would result in predicting a different outcome $\mathbf{c}'$, with $\mathbf{c} \neq \mathbf{c}'$. Although defeasible reasoning is not logically valid (the principle in itself is non-monotonic), it is rationally compelling and we therefore see it as an asset for explaining AI predictions in case of uncertainty over one or more variables.

The concept of defeasible reasoning has previously been used in the context of explaining Bayesian networks: defeasible rules were extracted from an alternative representation of the Bayesian network and subsequently used in an argumentation system to provide an account of the underlying reasoning that is easier to understand by the end user [27,35]. The resulting argumentative tree allows for showing the arguments that were invalidated due to lack of evidence.

We emphasise that defeasible reasoning is different from counterfactual reasoning: in the latter, all variables are observed and we look for hypothetical scenarios that alter our predictions, whereas with the former, we are interested in knowing how additional evidence will alter these predictions.

## 3. Relevance and MAP-independence

In this section we explicitly tie relevance to the concept of MAP-independence. We will first study theoretical properties of MAP-independence and subsequently discuss what the consequences are for finding sets of relevant nodes.

### 3.1. Relevance: definitions and properties

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. Kwisthout [17] 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 and focuses on the computational complexity of determining whether a given set $\mathbf{R} \subseteq \mathbf{S}$ is (ir)relevant for motivating a MAP-explanation. However, the question of how to determine the partition of $\mathbf{S}$ into relevant and irrelevant nodes is not addressed. Trying to facilitate finding such a partition, we set out to study properties of (ir)relevant nodes. First, we explicitly define (ir)relevance in the context of the MAP-explanation as follows.

**Definition 4.** Let $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ and $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ be as before. Then $\mathbf{R} \subseteq \mathbf{S}$ is *irrelevant* to $\mathbf{h}^*$ if $\mathbf{h}^*$ is MAP-independent of $\mathbf{R}$; otherwise $\mathbf{R}$ is *relevant* to $\mathbf{h}^*$. A set $\mathbf{R}$ that is (ir)relevant to $\mathbf{h}^*$ is called a(n) *(ir)relevance set*.

The terms irrelevant and MAP-independent, as well as relevant and MAP-dependent, will now be used interchangeably throughout this work.

We first consider the relation to computationally-irrelevant nodes such as d-separated and barren nodes and conclude that these can in fact be relevant according to Definition 1.

**Proposition 1.** *Let* $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ *and* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ *be as before. Let* $\mathbf{R} \subseteq \mathbf{S}$ *be relevant to* $\mathbf{h}^*$. *Then* $\mathbf{R}' \subseteq \mathbf{R}$ can be *computationally-irrelevant to* $\mathbf{H}$ *given* $\mathbf{e}$.

Proposition 1 poses a rather weak claim that can be substantiated through Example 1, below. In this and following examples we will exploit the following equality:

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

**Example 1.** Consider the Bayesian network in Fig. 2, with the following conditional probabilities tables (CPTs) specified for nodes $A$, $B$ and $C$ (complements omitted):

$$\Pr(a) = 0.55, \quad \Pr(c \mid a\,b) = 0.80, \quad \Pr(c \mid a\,\overline{b}) = 0.55,$$
$$\Pr(b) = 0.60, \quad \Pr(c \mid \overline{a}\,b) = 0.10, \quad \Pr(c \mid \overline{a}\,\overline{b}) = 0.50.$$

Let $\mathbf{E} = \emptyset$ and $\mathbf{H} = \{A\}$; note that in this case $\mathbf{h}^* = a$. Now, 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 $\mathbf{E}$. Indeed $B$ is irrelevant to $\mathbf{h}^*$, since $\Pr(a) = \Pr(a \mid b) = \Pr(a \mid \overline{b})$. However, both $\{C\}$ and $\{B, C\}$ are relevant to $\mathbf{h}^*$: e.g., $\Pr(\overline{a} \mid \overline{c}) = 0.67$ and although $\Pr(a \mid \overline{b}\,\overline{c}) = 0.52$, we have that $\Pr(\overline{a} \mid b\,\overline{c}) = 0.79$. We conclude that $B$ is irrelevant to $\mathbf{h}^*$, but becomes relevant in combination with node $C$. □

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

**Proposition 2.** *Let* $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ *and* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ *be as before, and let* $\mathbf{R}' \subseteq \mathbf{R} \subseteq \mathbf{S}$. *If* $\mathbf{R}'$ *is relevant to* $\mathbf{h}^*$ *then* $\mathbf{R}$ *is relevant to* $\mathbf{h}^*$.

**Proof.** We will prove the equivalent proposition which states that $\mathbf{h}^*$ is MAP-independent of $\mathbf{R}'$ if $\mathbf{h}^*$ is MAP-independent of $\mathbf{R}$. Suppose that $\mathbf{h}^*$ is MAP-independent of $\mathbf{R}$, then for all $\mathbf{r}' \in \Omega(\mathbf{R}')$ and all $\mathbf{r}'' \in \Omega(\mathbf{R} \setminus \mathbf{R}')$ we have $\arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h}\,\mathbf{r}'\,\mathbf{r}'' \mid \mathbf{e}) = \mathbf{h}^*$. By definition, this means that $\Pr(\mathbf{h}^*\,\mathbf{r}'\,\mathbf{r}'' \mid \mathbf{e}) > \Pr(\mathbf{h}\,\mathbf{r}'\,\mathbf{r}'' \mid \mathbf{e})$ for all $\mathbf{r}'$, $\mathbf{r}''$ and all $\mathbf{h} \neq \mathbf{h}^*$. The inequality then also holds for the sum over $\mathbf{r}''$: $\sum_{\mathbf{r}''} \Pr(\mathbf{h}^*\,\mathbf{r}'\,\mathbf{r}'' \mid \mathbf{e}) > \sum_{\mathbf{r}''} \Pr(\mathbf{h}\,\mathbf{r}'\,\mathbf{r}'' \mid \mathbf{e})$. As a result, if we marginalise over $\mathbf{R}''$, we have that $\arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h}\,\mathbf{r}' \mid \mathbf{e}) = \arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \sum_{\mathbf{r}''} \Pr(\mathbf{h}\,\mathbf{r}'\,\mathbf{r}'' \mid \mathbf{e}) = \mathbf{h}^*$ for all $\mathbf{r}' \in \Omega(\mathbf{R}')$. We conclude then that $\mathbf{h}^*$ is MAP-independent of $\mathbf{R}'$. □

In Example 1 we showed that a d-separated and hence computationally irrelevant node was irrelevant, but became relevant in combination with a relevant node. Irrelevance is not necessarily due to computational irrelevance, and the relevance of a set of nodes is not necessarily due to the relevant individual nodes in the set, as demonstrated in the following example.

**Example 2.** Reconsider the Bayesian network in Fig. 2, with the following probabilities specified for nodes $A$, $B$ and $C$ (complements omitted):

$$\Pr(a) = 0.80, \quad \Pr(c \mid a\,b) = 0.90, \quad \Pr(c \mid a\,\overline{b}) = 0.70,$$
$$\Pr(b) = 0.70, \quad \Pr(c \mid \overline{a}\,b) = 0.60, \quad \Pr(c \mid \overline{a}\,\overline{b}) = 0.20.$$

Let $\mathbf{E} = \emptyset$ and $\mathbf{H} = \{C\}$; we find $\Pr(c) = 0.768$, so $\mathbf{h}^* = c$. Note that both $A$ and $B$ are *computationally* relevant for $C$ given $\mathbf{E}$, yet only node $A$ is relevant to $\mathbf{h}^*$: $\Pr(c \mid a) = 0.84$, $\Pr(c \mid \overline{a}) = 0.48$, $\Pr(c \mid b) = 0.84$, and $\Pr(c \mid \overline{b}) = 0.6$. The set $\{A, B\}$ is clearly relevant to $\mathbf{h}^*$, since the specified probabilities $\Pr(c \mid A\,B)$ are not all above 0.5, meaning that some value assignment may alter $\mathbf{h}^*$. Moreover, the relevance of $\{A, B\}$ is not just due to the relevance of $A$, since this would imply that $\Pr(c \mid a\,B) \geq 0.5$ and $\Pr(c \mid \overline{a}\,B) < 0.5$ for all values of $B$, which is not the case. Finally note that $B$ is not irrelevant in combination with $A$ $(= \overline{a})$ and $A$ is in fact irrelevant in combination with $B$ $(= \overline{b})$. □

Whereas Proposition 2 entails that every subset of an irrelevant set is irrelevant, not every union of irrelevant sets is guaranteed to remain irrelevant.

**Proposition 3.** *Let* $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ *and* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ *be as before. Let* $\mathbf{R}, \mathbf{R}' \subseteq \mathbf{S}$ *both be irrelevant to* $\mathbf{h}^*$. *Then* $\mathbf{R} \cup \mathbf{R}'$ can be *relevant to* $\mathbf{h}^*$.

We again substantiate this claim through an example.

**Example 3.** Consider the Bayesian network in Fig. 2, with the following CPTs specified for nodes $A$, $B$, $C$ and $D$ (complements omitted):

$$\Pr(a) = 0.30, \quad \Pr(c \mid a\,b) = 0.50, \quad \Pr(c \mid a\,\overline{b}) = 0.70, \quad \Pr(d \mid c) = 0.25,$$
$$\Pr(b) = 0.50, \quad \Pr(c \mid \overline{a}\,b) = 0.40, \quad \Pr(c \mid \overline{a}\,\overline{b}) = 0.80, \quad \Pr(d \mid \overline{c}) = 0.70.$$

Let $\mathbf{E} = \{B\}$ with $\mathbf{e} = \overline{b}$, and let $\mathbf{H} = \{C\}$. Then $\mathbf{h}^* = c$, since $\Pr(c \mid \overline{b}) = 0.77$. Both $\{A\}$ and $\{D\}$ are irrelevant to $c$: $\Pr(c \mid \overline{a}\,\overline{b}) = 0.8$, $\Pr(c \mid a\,\overline{b}) = 0.7$, $\Pr(c \mid \overline{d}\,\overline{b}) = 0.89$, and $\Pr(c \mid d\,\overline{b}) = 0.54$. However, $\Pr(c \mid a\,d\,\overline{b}) = 0.45$ and therefore $\{A, D\}$ is relevant to $\mathbf{h}^* = c$. □

A similar example—where the MAP-explanation was MAP-independent of two individual nodes, but not of their combination—motivated the introduction of the concept of weak MAP-independence [18]. Weak MAP-independence is implied by MAP-independence.

**Proposition 4.** *Let* $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ *and* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ *be as before. If* $\mathbf{h}^*$ *is* MAP*-independent of* $\mathbf{R} \subseteq \mathbf{S}$ *then* $\mathbf{h}^*$ *is weakly* MAP*-independent of* $\mathbf{R}$.

**Proof.** This follows directly by applying Proposition 2 to $\mathbf{R}' = \{R\}$ for each $R \in \mathbf{R}$. □

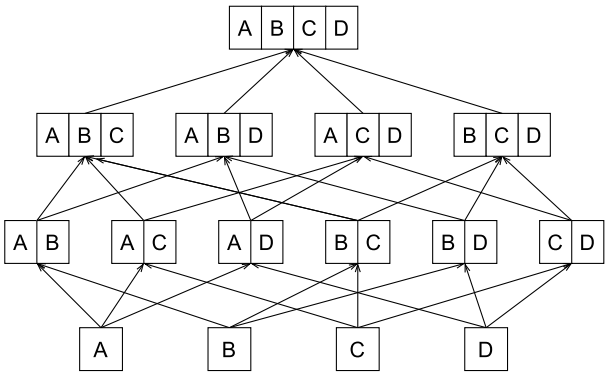


**Fig. 3.** Search space of all potential (ir)relevant node sets, $\mathcal{P}(\mathbf{S})$, for $\mathbf{S} = \{A, B, C, D\}$.

The reverse implication is only guaranteed for singleton sets $R \in \mathbf{S}$.

**Proposition 5.** *Let* $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ *and* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ *be as before. If the* MAP*-explanation* $\mathbf{h}^*$ *is weakly* MAP*-independent of* $R \in \mathbf{S}$ *then* $\mathbf{h}^*$ *is* MAP*-independent of* $R$.

**Proof.** This follows immediately from Definitions 1 and 2. □

From Propositions 2 and 3 and the above examples 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 in [17], 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 $\mathcal{P}(\mathbf{S})$. Note that the empty set can be excluded, since MAP-independence is a trivial problem for $\mathbf{R} = \emptyset$.

### *3.2. Partitioning the powerset*

The observation that we need to partition the powerset $\mathcal{P}(\mathbf{S})$ unfortunately makes the problem of finding relevance sets only harder, as the search space becomes exponentially large (see Fig. 3). We therefore analyse this partition further. To this end, we first introduce two new concepts related to MAP-(in)dependence or (ir)relevance.

**Definition 5.** Let $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ and $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ be as before and let $\mathbf{h}^*$ be MAP-independent of $\mathbf{R} \subseteq \mathbf{S}$. If either $\mathbf{R} = \mathbf{S}$ or $\mathbf{h}^*$ is MAP-*de*pendent of $\mathbf{R} \cup \{S'\}$, $\forall\, S' \in \mathbf{S} \setminus \mathbf{R}$, then $\mathbf{h}^*$ is *maximal* MAP*-independent* of $\mathbf{R}$.

The above definition suffices to exclude all supersets of $\mathbf{R}$ to be MAP-independent, which justifies denoting $\mathbf{R}$ as a maximal set. Note that a maximal MAP-independent set is not necessarily a maxim*um* MAP-independent set (see Definition 3).

**Proposition 6.** *Let* $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ *and* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ *be as before. If* $\mathbf{h}^*$ *is maximal* MAP*-independent of* $\mathbf{R} \subseteq \mathbf{S}$ *then* $\mathbf{h}^*$ *is* MAP*-dependent of any* $\mathbf{R}' \supset \mathbf{R}$, $\mathbf{R}' \subseteq \mathbf{S}$.

**Proof.** Suppose there does exist a $\mathbf{R}' \supset \mathbf{R}$, $\mathbf{R}' \subseteq \mathbf{S}$, such that $\mathbf{h}^*$ is MAP-independent of $\mathbf{R}'$. Then, by Proposition 2, $\mathbf{h}^*$ is MAP-independent of every subset of $\mathbf{R}'$, including every subset $\mathbf{R} \cup \{S'\}$ for $S' \in \mathbf{R}' \cap (\mathbf{S} \setminus \mathbf{R})$. This contradicts our assumption that $\mathbf{R}$ is a maximal MAP-independent set. □

Analogous to maximal MAP-independence, we define minimal MAP-*de*pendence.

**Definition 6.** Let $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ and $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ be as before and let $\mathbf{h}^*$ be MAP-dependent of $\mathbf{R} \subseteq \mathbf{S}$. If either $\mathbf{R}$ is a singleton or $\mathbf{h}^*$ is MAP-independent of $\mathbf{R}'$, $\forall \mathbf{R}' \subset \mathbf{R}$, then $\mathbf{h}^*$ is *minimal* MAP*-dependent* of $\mathbf{R}$.

We now introduce some further notations:

$\mathcal{P}^+(\mathbf{S})$**:** subset of $\mathcal{P}(\mathbf{S})$ containing all relevant (MAP-dependent) subsets of $\mathbf{S}$;
$\mathcal{P}^-(\mathbf{S})$**:** subset of $\mathcal{P}(\mathbf{S})$ containing all irrelevant (MAP-independent) subsets of $\mathbf{S}$;
$\mathcal{P}^{m+}(\mathbf{S})$**:** subset of $\mathcal{P}(\mathbf{S})$ containing all minimal relevant subsets of $\mathbf{S}$;
$\mathcal{P}^{M-}(\mathbf{S})$**:** subset of $\mathcal{P}(\mathbf{S})$ containing all maximal irrelevant subsets of $\mathbf{S}$.

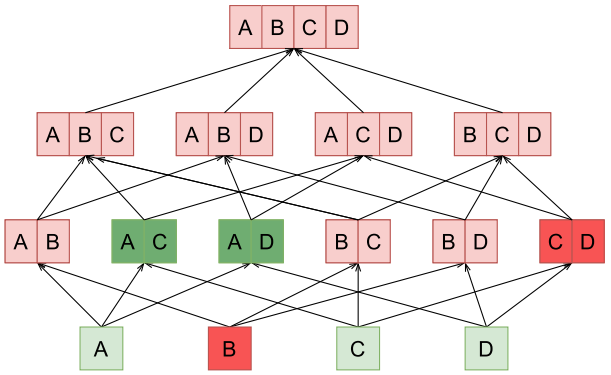


**Fig. 4.** Search space $\mathcal{P}(\mathbf{S})$, for $\mathbf{S} = \{A, B, C, D\}$. In green, MAP-independent (irrelevant) sets; in red MAP-dependent (relevant) sets. In darker green, maximal MAP-independent sets; in darker red, minimal MAP-dependent sets. A graphical explanation of Definitions 5 and 6 can be given as well using the search space graph: a set of nodes is maximal MAP-independent if it has no MAP-independent children, whereas a set is minimal MAP-dependent if it has no MAP-dependent parents. In the search space graph, children are immediately above the parents and connected by arcs.

Some interesting properties to characterise $\mathcal{P}^{+}(\mathbf{S})$ and $\mathcal{P}^{-}(\mathbf{S})$ can be derived from Definitions 1, 2, 5 and 6. We first observe that any (strict) superset of a maximal MAP-independent set is by definition relevant. Moreover, recall that any subset of an irrelevant set is irrelevant (Proposition 2). As a result, we find the following relation between $\mathcal{P}^{+}(\mathbf{S})$, $\mathcal{P}^{-}(\mathbf{S})$ and maximal MAP-independence.

**Proposition 7.** *Let* $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ *and* $\mathbf{h}^{*} = \text{MAP}(\mathbf{H}, \mathbf{e})$ *be as before. Then the sets* $\mathcal{P}^{+}(\mathbf{S})$ *and* $\mathcal{P}^{-}(\mathbf{S})$ *can be derived from* $\mathcal{P}^{M-}(\mathbf{S})$.

**Proof.** Initialise both $\mathcal{P}^{+}(\mathbf{S})$ and $\mathcal{P}^{-}(\mathbf{S})$ as empty. Then, $\forall \mathbf{S}' \in \mathcal{P}(\mathbf{S})$, insert $\mathbf{S}'$ in $\mathcal{P}^{-}(\mathbf{S})$ if $\mathbf{S}' \subseteq \mathbf{M}'$, $\forall \mathbf{M}' \in \mathcal{P}^{M-}(\mathbf{S})$. Else, insert $\mathbf{S}'$ into $\mathcal{P}^{+}(\mathbf{S})$.

By the logical opposite of Proposition 2, we know that any subset of an irrelevant set will also be irrelevant, and thus the "if" part of the if-else block is justified. We know that no other set in $\mathcal{P}(\mathbf{S})$ is irrelevant, otherwise it should have been contained in any set of $\mathcal{P}^{M-}(\mathbf{S})$, according to its own definition (see Definition 5). □

This compacts the entirety of the set of exponential size $\mathcal{P}(\mathbf{S})$ into a strongly reduced set $\mathcal{P}^{M-}(\mathbf{S})$ without losing any information. A natural question that arises here is whether we can also represent $\mathcal{P}(\mathbf{S})$ by exploiting minimal MAP-dependence. In this case we again recall that every superset of a relevant set is relevant, and note that any (strict) subset of a minimal MAP-dependent set is by definition irrelevant, we then find the following result:

**Proposition 8.** *Let* $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ *and* $\mathbf{h}^{*} = \text{MAP}(\mathbf{H}, \mathbf{e})$ *be as before. Then the sets* $\mathcal{P}^{+}(\mathbf{S})$ *and* $\mathcal{P}^{-}(\mathbf{S})$ *can be derived from* $\mathcal{P}^{m+}(\mathbf{S})$.

**Proof.** Initialise both $\mathcal{P}^{+}(\mathbf{S})$ and $\mathcal{P}^{-}(\mathbf{S})$ as empty. Then, $\forall \mathbf{S}' \in \mathcal{P}(\mathbf{S})$, insert $\mathbf{S}'$ in $\mathcal{P}^{+}(\mathbf{S})$ if $\mathbf{M}' \subseteq \mathbf{S}'$, $\forall \mathbf{M}' \in \mathcal{P}^{m+}(\mathbf{S})$. Else, insert $\mathbf{S}'$ into $\mathcal{P}^{-}(\mathbf{S})$.

By Proposition 2, we know that any superset of a relevant set will also be relevant, justifying the "if" part of the if-else block. We know that no other set in $\mathcal{P}(\mathbf{S})$ is relevant, as the definition of the set $\mathcal{P}^{m+}(\mathbf{S})$ would be incomplete (see Definition 6). □

While Propositions 7 and 8 are not an advance in terms of speed up, they are significant in terms of memory needed and also explainability due to the enhanced simplicity of sets $\mathcal{P}^{M-}(\mathbf{S})$ and $\mathcal{P}^{m+}(\mathbf{S})$. Specifically, we avoid showing to the user any set $\mathbf{R}$ such that either $\exists \mathbf{M}' \in \mathcal{P}^{M-}(\mathbf{S}) : \mathbf{R} \subset \mathbf{M}'$ or $\exists \mathbf{M}' \in \mathcal{P}^{m+}(\mathbf{S}) : \mathbf{M}' \subset \mathbf{R}$. Such simplification can be visualised in Fig. 4.

## 4. Finding relevance sets

A straightforward approach to actually finding a partition of $\mathcal{P}(\mathbf{S})$ into relevant and irrelevant subsets would be to test the relevance of each subset $\mathbf{S}_i$ of $\mathbf{S}$ in a brute-force way. This approach would amount to performing $\sum_{i=1}^{2^{|\mathbf{S}|}} |\Omega(\mathbf{S}_i)|$ MAP-checkings. Since such an approach does not necessarily exploit any Bayesian network properties, we note that it can be used for finding sets of (ir)relevant variables from other type of models such as those found in domains of predictive modelling (see the end of Section 2.2) where uncertainty is considered and can be handled (for example, using an imputer for the unobserved variables). However, such cases might be even harder to deal with, since we do not have the advantage provided by the Bayesian network of explicitly and compactly modelling probability distributions. While these cases are interesting to give a baseline comparison to MAP-independence in Bayesian networks, studying them in detail is beyond the scope of this work.

In either case, the brute-force approach is clearly inefficient, so alternative approaches are called for. Since the relevance sets are to be used for explanation purposes and explanations should be as simple as possible [24], we aim for small sets of relevant nodes. Moreover, the computational cost of establishing the relevance of a given set $\mathbf{R}$ depends heavily on its size [17]. 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 (as singletons) 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 most 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 concluded that d-separated nodes cannot be relevant in isolation; as a result, they can be pruned from the network [3]:

**Proposition 9.** *Let* $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ *and* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ *be as before. 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, because there are no active (unblocked) chains between $R$ and $H_i \in \mathbf{H}$ given $\mathbf{E}$. □

Individual barren nodes are also computationally irrelevant, but can nevertheless 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 Bayes classifier: every unobserved input node is a barren node with respect to the output node $C$. 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 do *not* disregard barren nodes with respect to $\mathbf{H}$ and $\mathbf{E}$ as possible relevant nodes.

A naive approach to establish all relevant singletons would again be to iterate over all $R \in \mathbf{R}$ and $r \in \Omega(R)$ and compute the MAP-explanation for each evidence configuration $(r\,\mathbf{e})$. This brute-force approach would require at most $|\mathbf{R}| \cdot \max_{R \in \mathbf{R}} |\Omega(R)|$ MAP-checkings (checking if $\arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h}\, r \mid \mathbf{e}) = \mathbf{h}^*$ for a given $r \in \Omega(R)$). Instead, we propose to exploit the Bayesian network representation by iterating over all $\mathbf{h} \in \Omega(\mathbf{H})$ and computing posterior distributions $\Pr(R \mid \mathbf{h}\,\mathbf{e})$ using standard (join-tree) inference. We thereby exploit the following property:

**Proposition 10.** *Let* $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ *and* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ *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*

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

**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

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

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, i.e., after at most $|\Omega(\mathbf{H})| - 1$ full network propagations, it is guaranteed to have identified the set $\mathcal{R}$ of all nodes that constitute relevant singletons. The pseudocode for this method is provided by Algorithm 1.

The algorithm also serves to establish all subsets of $\mathbf{S}$ to which $\mathbf{h}^*$ is weakly MAP-independent.

**Proposition 11.** *Let* $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ *and* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ *be as before. Let* $\mathcal{R}$ *be the set of relevant singletons returned by Algorithm 1. Then* $\mathbf{h}^*$ *is weakly* MAP*-independent of any* $\mathbf{R} \subseteq \mathbf{S} \setminus \mathcal{R}$.

**Proof.** This follows directly from the observation that $\mathbf{S} \setminus \mathcal{R}$ contains all nodes that are individually MAP-independent of $\mathbf{h}^*$ and Proposition 4. □

**Algorithm 1:** Computing relevant singletons for $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$.

**Input :** (pruned) Bayesian network $\mathcal{B}$, $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$, $\mathbf{c}_i = \log[\Pr(\mathbf{h}_i\,\mathbf{e})/\Pr(\mathbf{h}^*\,\mathbf{e})]$ 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**
**end**
**return** $\mathcal{R}$

### 4.2. Extending to larger relevance sets

Recall from Proposition 2 that any superset of relevant singletons for $\mathbf{h}^*$ is also a relevance set for $\mathbf{h}^*$. Given all relevant singletons, therefore, allows us to immediately establish many, yet not all, larger relevance sets by using Proposition 2 and labelling as relevant every superset of the singletons from $\mathcal{R}$ returned by Algorithm 1. However, for the purpose of explaining the robustness of a MAP-explanation, relevance sets should preferably not include nodes that do not actually contribute to the relevance and this idea falls quite short.

We propose an algorithm that exploits the properties of Section 3 to find the powerset partition into relevant and irrelevant sets, $\mathcal{P}^+(\mathbf{S})$ and $\mathcal{P}^-(\mathbf{S})$ respectively. Pseudocode is provided by Algorithm 2.

#### 4.2.1. Suggestions to prune the exhaustive search space

We first propose suggestions that will be used alongside the propositions of Section 3 to traverse and prune the search space. These prunes are not heuristic, i.e., the whole powerset is correctly partitioned in relevant and irrelevant subsets. They only avoid unnecessary or redundant computations.

Recall from Proposition 3 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 our attention to relevant combinations of irrelevant sets only (i.e., those that fulfil Proposition 3). Therefore, it is reasonable to only check for the (ir)relevance of unions of irrelevant sets of nodes. An option is to start by evaluating every pair of irrelevant nodes and expanding the sets as long as the combination remains irrelevant.

**Suggestion 1.** *Let $\mathcal{R}$ be the set of relevant singletons for $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ as returned by Algorithm 1. Initialise $\mathcal{I} = \mathbf{S} \setminus \mathcal{R}$ to the set of irrelevant singletons. Check each pair of nodes from $\mathcal{I}$ for their relevance. If the pair is relevant to $\mathbf{h}^*$ then we add it to $\mathcal{R}$ and stop investigating its supersets, otherwise we increase the subsets of $\mathcal{I}$ by one and investigate its relevance.*

The above suggestion implies that the search space, as depicted in Fig. 3, is built only from irrelevant singletons. The search space is now exponential in the number of irrelevant singletons. However, to alleviate the computational burden that such search implies, we refrain from investigating supersets of nodes that include relevant subsets, as we already know that such supersets will be relevant by Proposition 2. This allows us to safely prune the search space, and thus we will only work with the union of irrelevant nodes. To further reduce the number of irrelevant sets to evaluate, again, d-separation may be employed.

**Proposition 12.** *Let $\mathbf{H}, \mathbf{E}, \mathbf{S}$, and $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ be as before, and let $\mathbf{U}, \mathbf{W} \subset \mathbf{S}$. 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 $\mathbf{h}^*$.*

**Proof.** Assume $\mathbf{U}$ is irrelevant to $\mathbf{h}^*$, that is $\mathbf{h}^* = \arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h}\,\mathbf{u} \mid \mathbf{e}) = \arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h} \mid \mathbf{u}\,\mathbf{e})$ for all $\mathbf{u} \in \Omega(\mathbf{U})$. Since d-separation implies independence, we have that for all $\mathbf{w} \in \Omega(\mathbf{W})$

$$\arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h} \mid \mathbf{u}\,\mathbf{e}) = \arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h} \mid \mathbf{w}\,\mathbf{u}\,\mathbf{e}) = \arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h}\,\mathbf{w}\,\mathbf{u} \mid \mathbf{e})$$

But then also $\arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \sum_{\mathbf{u}} \Pr(\mathbf{h}\,\mathbf{w}\,\mathbf{u} \mid \mathbf{e}) = \arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h}\,\mathbf{w} \mid \mathbf{e}) = \mathbf{h}^*$ (same principle as in the proof for Proposition 2). We conclude that $\mathbf{W}$ is irrelevant to $\mathbf{h}^*$. □

A set that d-separates a node from every other node in the DAG is its *Markov blanket* MB(), consisting (assuming the faithfulness property in the DAG) of the node's parents, children, and co-parents (also known as spouses)[3] [25]. If the Markov blanket of a hypothesis node $H$ was to be irrelevant to the most likely value $h^*$ of the node, given the evidence, then the whole $\mathbf{S}$ is irrelevant to $h^*$. Establishing the (ir)relevance of the Markov blanket of $H$ can be done by inspecting the CPTs specified locally for $H$ and its children.

**Proposition 13.** *Let* $\mathbf{H} = \{H\}$ *and* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ *be as before and let* $Ch(H)$ *denote the set of all the children of* $H$. *Then* $\mathbf{M} = \text{MB}(H)$ *(its Markov blanket) is irrelevant to* $h^*$ *iff* $\forall \mathbf{m} \in \Omega(\mathbf{M})$ *and* $\forall h \neq h^* \in \Omega(H)$*:*

$$\Pr(h^* \mathbf{m} \mid \mathbf{e}) > 0 \text{ and } \sum_{C \in Ch(H)} \log \frac{\Pr(c \mid h \boldsymbol{\pi}_{C \setminus H})}{\Pr(c \mid h^* \boldsymbol{\pi}_{C \setminus H})} + \log \frac{\Pr(h \mid \boldsymbol{\pi}_H)}{\Pr(h^* \mid \boldsymbol{\pi}_H)} \leq 0,$$

*where the configurations* $c \in \Omega(C)$, $\boldsymbol{\pi}_{C \setminus H} \in \Omega(\boldsymbol{\pi}(C) \setminus \{H\})$ *and* $\boldsymbol{\pi}_H \in \Omega(\boldsymbol{\pi}(H))$ *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* $\mathbf{e}$.

**Proof.** By definition, $\text{MB}(H)$ 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 also employ Proposition 13 in the case $\mathbf{H}$ contains more than one node.

**Proposition 14.** *Let* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ *be as before. Let* $\mathbf{H} = \{H_1, \ldots, H_n\}$, $n > 1$, *and let* $h_i^* = \text{MAP}(H_i, \mathbf{e})$, $h_i^* \in \Omega(H_i)$, *for each* $H_i \in \mathbf{H}$. *If for each* $H_i$, $\text{MB}(H_i)$ *is irrelevant to* $h_i^*$ *then* $\mathbf{h}^* = h_1^* \ldots h_n^*$ *and* $\cup_{i=1}^n \text{MB}(H_i)$ *is irrelevant to* $\mathbf{h}^*$.

**Proof.** It follows directly from the conditional independence properties given the Markov blankets. □

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

**Example 4.** Consider the Insurance Bayesian network (Fig. 5) with evidence for $|\mathbf{E}| = 5$ nodes: Age = `senior`, SocioEcon = `uppermiddle`, DrivHist = `many`, HomeBase = `city`, MakeModel = `sportscar`. We examined the relevance of the Markov blanket for each of the 22 other nodes in the graph, which will play the role of $H$ given the evidence. One node, GoodStudent, 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, their Markov blanket is irrelevant to their most likely value. Note that both of these nodes have only a single node in their Markov blanket. □

In case the entirety of the Markov blanket of $H$ is not irrelevant, we could determine the irrelevant subset $\mathbf{R}'$ and subsequently add the Markov blankets of the nodes $\text{MB}(H) \setminus \mathbf{R}'$ 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 blankets, Proposition 14 no longer applies and, as will be shown in a more detailed experiment in Section 6.4, we cannot in general aggregate the results for multiple hypothesis nodes to draw conclusions about (ir)relevance to a most likely combination of hypotheses.

We conclude that starting from irrelevant Markov blankets as described above does not seem a very promising alternative to Algorithm 2, but we can exploit the Markov blankets to some extent within the algorithm.

**Suggestion 2.** *Let* $\mathbf{H}$ *be as before and let* $\mathcal{I}$ *and* $\mathcal{P}$ *be as initialised in Algorithm 2. Upon ordering the subsets of* $\mathcal{P}$ *in line 3 by size, order the subsets* $\mathbf{P}_i \in \mathcal{P}$ *of equal size such that those* $\mathbf{P}_i \subseteq \text{MB}(H_i)$, $H_i \in \mathbf{H}$, *come first.*

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

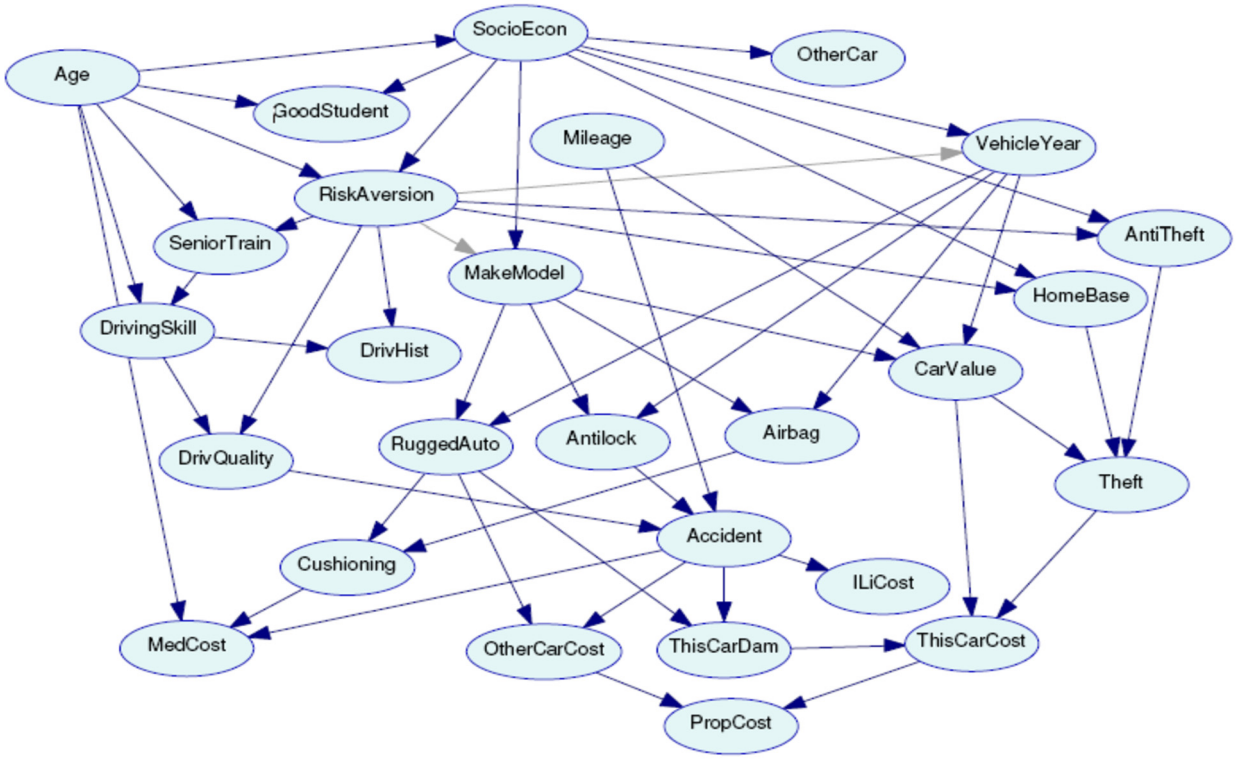


**Fig. 5.** The graph of the Insurance Bayesian network [5].

**Algorithm 2:** Computing all (ir)relevance sets for $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$.

**Input :** Bayesian network $\mathcal{B}$ with nodes $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$, $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$
**Output:** set $\mathcal{R}$ of minimal relevant subsets, set $\mathcal{I}$ of maximal irrelevant subsets
1 $\mathcal{R} \leftarrow$ compute relevant singletons (Algorithm 1); $\mathcal{I} \leftarrow \mathbf{S} \setminus \mathcal{R}$
2 $\mathcal{P} \leftarrow powerset(\mathcal{I})$ # Candidates to relevant nodes sets, Suggestion 1
3 Order $\mathcal{P}$ by size (small to large sets). Inside the same size, prioritise sets with nodes in the MB of $H \in \mathbf{H}$ # Suggestion 2
4 **forall** $\mathbf{P}_i \in \mathcal{P}$ **do**
5 **if** *not* MAP-ind($\mathbf{P}_i, \mathbf{h}^*, \mathbf{e}$) **then**
6 $\mathcal{R} \leftarrow \mathcal{R} \cup \mathbf{P}_i$
7 $\mathcal{P} \leftarrow \{\mathbf{R}' \in \mathcal{P} | \mathbf{P}_i \not\subset \mathbf{R}'\}$ # Prune supersets, Proposition 2
8 **end**
9 **else**
10 $\mathcal{I} \leftarrow \mathcal{I} \cup \mathbf{P}_i$
11 Prune sets d-separated from $\mathbf{H}$ given $\mathcal{I}$ # Proposition 12
12 **end**
13 **end**
14 Eliminate from $\mathcal{I}$ all subsets of other sets # Propositions 6 and 7
15 **return** $\mathcal{R}, \mathcal{I}$

If we have only a single hypothesis node, then the (ir)relevance of Markov blanket nodes can be established quite efficiently. However, in this case each MAP computation requires just a single (full) network propagation, which already hugely simplifies the problem of finding relevance sets in the first place. For multiple hypothesis nodes, the local computations in the Markov blanket may not necessarily be exploited, so the overhead of establishing the ordering as suggested may not be worthwhile. Still, it is included in the algorithm, as it may pose some benefits.

Finally, when the exhaustive search reaches too large sets to evaluate, we can decide to prematurely stop the search. In that case, the sets output by the algorithm are no longer guaranteed to be the complete $\mathcal{P}^{m+}$ and $\mathcal{P}^{M-}$.

**Suggestion 3.** *We can stop the search process any time we have found enough relevance sets of sufficient size for explanation purposes (see the discussion in Section 5).*

*4.2.2. Description of the exhaustive algorithm*

The exhaustive algorithm adopts a breadth-first search approach and its pseudocode is presented in Algorithm 2. Line 7

uses Proposition 2 to remove from the search space the sets that are known to be relevant, therefore only considering the union of irrelevant sets for evaluation (Proposition 3). Line 11 applies a pruning based on Proposition 12 that is even further optimised by prioritising nodes in the Markov blanket of $H \in \mathbf{H}$ (Suggestion 2). In addition, lines 7 and 14 ensure that the sets yielded are actually $\mathcal{P}^{M-}(\mathbf{S})$ and $\mathcal{P}^{m+}(\mathbf{S})$ (see Propositions 7 and 8): by removing all relevant supersets from $\mathcal{P}$ in line 7, these will not be added to the set $\mathcal{R}$ and hence $\mathcal{R}$ includes all minimal relevant subsets; the reduction step in line 14 ensures that $\mathcal{I}$ only includes the maximal irrelevant subsets. Although Suggestion 3 is not included in the pseudocode, it has been implemented in the Python code used for experimentation and lets us show the user the maximum size of irrelevant sets to find.

The sets output by the algorithm provide an elegant and efficient representation of (ir)relevance sets, thereby enhancing the overall explainability. As long as explanations focus only on the *nodes* in the relevance sets, returning minimal relevance sets is perfectly fine, since smaller sets provide simpler explanations. Optionally, it is possible to include the specific observations for these nodes that render them relevant. However, then we may want to consider supersets too: it could be the case that node $R_1$ is irrelevant to $\mathbf{h}^*$, node $R_2$ is relevant because $\bar{r}_2 \in \Omega(R_2)$ results in a MAP-explanation $\mathbf{h} \neq \mathbf{h}^*$, and that the combination is relevant because $(r_1\, r_2) \in \Omega(R_1) \times \Omega(R_2)$ together result in yet another MAP explanation.

### 4.3. Suggestions for improving exhaustive search results

Algorithm 2 returns the smallest possible sets of (ir)relevant nodes that carry all information for explaining the robustness of a MAP-explanation. In this section we explore further options for simplifying the output or the algorithm as a whole.

**Suggestion 4.** *Prior to establishing (ir)relevance of any size sets for* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$*, prune* $\mathbf{S}$ *by removing all nodes d-separated from* $\mathbf{H}$ *by* $\mathbf{E}$*.*

To motivate this pruning step, 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. Even when $V$ is irrelevant, if it unblocks a path between $D$ and $\mathbf{H}$, the set $\{V, D\}$ can be relevant by Proposition 2. This suggests that we should no longer disregard initially d-separated nodes when searching for larger relevance sets, as doing that will imply potentially hiding to the user (minimal) MAP-dependent sets that can alter the explanation $\mathbf{h}^*$. 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.

Even if all d-separated nodes suggested above are pruned (line 12), d-separation can be a useful tool for further 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 5.** *Let* $\mathbf{R}$ *be a relevance set for* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$*. 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* $\mathbf{R}$*.*

The above suggestion can be applied to the minimal relevant subsets in a post-processing step, getting an even more reduced version that considers properties of the DAG rather than just membership in other sets. Whether or not this is worth the effort will depend greatly on the structure of the Bayesian network and the size of the relevance sets in $\mathcal{P}^{m+}$.

We note that relevant nodes need to be observed in order to actually affect the MAP-explanation. Therefore, we could choose to evaluate only observable nodes for their relevance. Whether or not supplementary nodes are actually observable is of course domain-dependent, and that is why the following suggestion has not been included in our domain-agnostic algorithm.

**Suggestion 6.** *Let* $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ *and* $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$ *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.*

The suggestions in this section all have their benefits and drawbacks, as discussed. Rather than adapting Algorithm 2 to incorporate the suggestions, we decided to compare our approach to heuristic alternatives.

### 4.4. Heuristic alternatives

While the pruning applied in Algorithm 2 will likely speed up the run time, the search space is still exponential and we run into the problem of computing several MAP-independencies for large sets $\mathbf{R}$, which are notably slow to compute.

**Algorithm 3:** Approximate MAP-independence with FSS.

**Input :** Dataset $D(\mathbf{I}, \mathbf{H} \cup \mathbf{E} \cup \mathbf{S})$, evidence $\mathbf{e}$ and an FSS algorithm `FSS()`
**Output:** Set $\mathcal{R}$, set of all relevant variables
1 $D' \leftarrow D(\{\mathbf{i} \in \mathbf{I} | \mathbf{e} \subset \mathbf{i}\}, \mathbf{H} \cup \mathbf{S})$ # Steps (1) and (2)
2 $\mathcal{R} \leftarrow \texttt{FSS}(D')$ # Step (3)
3 **return** $\mathcal{R}$

#### 4.4.1. FSS to approximate MAP-independence

In many scenarios, we may be interested in knowing only the largest set(s) of irrelevant nodes from the supplementary nodes $\mathbf{S}$, instead of the full sets $\mathcal{P}^+(\mathbf{S})$ and $\mathcal{P}^-(\mathbf{S})$. If this is the case, the similarity of MAP-independence and the problem of FSS becomes more apparent: try to find the largest subset of irrelevant and redundant features. However, there are two key differences:

1. FSS can be seen as a global problem, since we try to find irrelevances that apply for every prediction, whereas MAP-independence is evidence-specific, restricting the problem to a more local scope.
2. Even mending the first gap, the goal of FSS is normally to increase a classifier performance, whereas in MAP-independence we care about whether a certain prediction/explanation $\mathbf{h}^*$ remains the same. While MAP-independence can be approximated using a FSS approach, the fundamental problem remains different and thus FSS shall be adapted.

Before formalising the idea, we introduce some notation. Let $D(\mathbf{I}, \mathbf{H} \cup \mathbf{E} \cup \mathbf{S})$ be a dataset, where $\mathbf{I}$ are the instances and $\mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$ the set of features (they are presented as a union to make the connection with MAP-independence in Bayesian networks easier).

It is possible to approximate the MAP-independence problem using FSS by "localizing" the dataset and focusing on certain instances such that the value assignment for the attributes $\mathbf{E}$ is $\mathbf{e}$. To do that, we need to (1) remove from the dataset the instances whose value-assignment for $\mathbf{E}$ is not $\mathbf{e}$, (2) then remove the features $\mathbf{E}$ and (3) finally apply the desired FSS algorithm to get the set of (ir)relevant nodes. The procedure is formalised in Algorithm 3.

Again, we would like to highlight that in theory this approach would be possible without a Bayesian network. However, while normally in step (1) of the aforementioned procedure the size of the dataset would be dramatically crippled and, in turn, the efficacy of the FSS techniques' would be compromised due to lack of samples, the Bayesian network allows us to generate more for the dataset.

Over the pre-processed dataset, we can apply a variety of FSS techniques, such as filter and wrapper approaches. Using filters is tempting, as they drastically reduce the computation time of looking for the largest set of irrelevant features: features are selected based on some score, typically mutual information, computed from the data. As such, these techniques are completely disconnected from the model for which the features are selected, and as a result we might run into the problem of obtaining a subset of supplementary nodes that the FSS algorithm identifies as irrelevant, yet in the Bayesian network is relevant. This scenario has been observed in some preliminary experiments. This problem could potentially be addressed using the concept of MAP-independence *strength* (see Section 5), that aims to measure *how* relevant a subset $\mathbf{R}$ of nodes is by computing the probability of obtaining any value-assignment for $\mathbf{R}$ that modifies the MAP-explanation $\mathbf{h}^*$. Using such a degree of relevance in the induced subset as a score, filter techniques can actually be used. Still, we would be labelling a set that is relevant as irrelevant and, in this work, we rather focus on obtaining faithful solutions.

#### 4.4.2. Foundations of a greedy search

Wrapper techniques, a priori, are not much more promising either: selecting a subset of features, then training a new Bayesian network and computing accuracy can actually be slower than computing MAP-independence. Wrapper approaches, however, use heuristics for navigating through the search space (greedy heuristics, genetic algorithms,...), and these heuristics are worth investigating in the context of MAP-independence.

We decided then to implement a hill climbing algorithm that iteratively adds a node to a subset of irrelevant nodes such that the resulting set remains irrelevant. The reason for selecting a simple hill climber rather than a more sophisticated heuristic algorithm is that this simple procedure may already perform similarly to our exhaustive approach in terms of quality of results. If this turns out not to be the case in our experiments, we can move towards more complex approaches in the future.

The first problem of using a heuristic algorithm for MAP-independence is how to guide the search. Testing MAP-independence for a subset of nodes gives a binary answer (relevant or irrelevant), which makes guiding the search impractical. Instead, we propose again to use some degree of relevance, where we take inspiration from a method that was proposed for approximating MAP-independence in Bayesian networks with continuous variables [36]. Rather than comparing the modes of $\Pr(\mathbf{H} \mid \mathbf{e})$ and $\Pr(\mathbf{H} \mid \mathbf{r}\mathbf{e})$, $\forall \mathbf{r} \in \Omega(\mathbf{R})$, the dissimilarity between the distributions $\Pr(\mathbf{H} \mid \mathbf{e})$ and $\Pr(\mathbf{H} \mid \mathbf{r}\mathbf{e})$ is captured using a divergence measure, under the hypothesis that higher dissimilarity is synonymous with a higher probability of the node set $\mathbf{R}$ being relevant (i.e. MAP-dependent). In this case, we use the maximum Jensen-Shannon divergence (JSD) [21] between $\Pr(\mathbf{H} \mid \mathbf{e})$ and $\Pr(\mathbf{H} \mid \mathbf{r}\mathbf{e})$, $\forall \mathbf{r} \in \mathbf{R}$, formally: $\max_{\mathbf{r} \in \mathbf{R}} \mathrm{JSD}(\Pr(\mathbf{H} \mid \mathbf{e}) \,||\, \Pr(\mathbf{H} \mid \mathbf{r}\mathbf{e}))$.

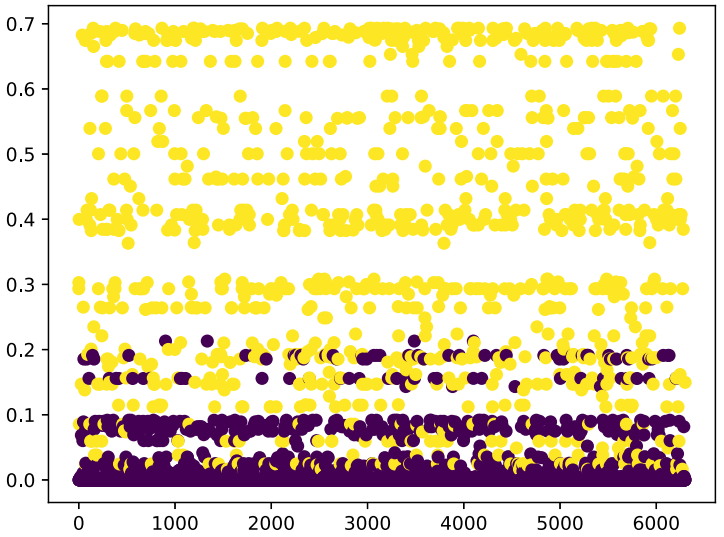


**Fig. 6.** Scatter plot, where the y-axis indicates $\max_{\mathbf{r}\in\mathbf{R}} \mathrm{JSD}(\Pr(H \mid \mathbf{e}) \,||\, \Pr(H \mid \mathbf{re}))$ for a sample of generated sets **R** from sizes between 1 and 6 to be evaluated. Every set **R** is distributed along the x-axis. Yellow dots mark sets that are relevant, whereas purple sets are irrelevant.

#### 4.4.3. Testing the hypothesis

We study the dependence between relevance and JSD empirically by modelling it as a problem with two variables: (1) *Rel*, a binary variable representing the (ir)relevance of a subset **R** of the set of supplementary nodes **S**, and (2) $JSD$, a continuous variable that models $\max_{\mathbf{r}\in\mathbf{R}} \mathrm{JSD}(\Pr(\mathbf{H} \mid \mathbf{e}) \,||\, \Pr(\mathbf{H} \mid \mathbf{re}))$. Then we use the Kolmogorov-Smirnov (KS) test [23] to prove that the densities $f(JSD \mid Rel = True)$ and $f(JSD \mid Rel = False)$ are statistically different (if there was no relation between them, the densities should be similar). For our experiment, we first generate 100 scenarios in which a single node $H$ is our hypothesis node, we have 6 supplementary nodes **S** and the rest of nodes are evidence nodes **E**. In each scenario, the actual nodes belonging to each subset are randomly chosen, as well as the value assignment of the evidence **e**, which is sampled using the Insurance Bayesian network (see Fig. 5). For each scenario, we will check MAP-independence and compute the maximum value of the JSD for every subset $\mathbf{R} \subseteq \mathbf{S}$. In Fig. 6, the results of our experiments are shown graphically.

In the figure, we can see that although correlation is not perfect, higher maximum values of JSD usually entail a higher probability of the node being relevant, and low maximum JSD values correspond to a higher probability of the node being irrelevant. When comparing $f(JSD \mid Rel = True)$ and $f(JSD \mid Rel = False)$, we get that the null hypothesis (the samples come from the same distribution) is rejected with a p-value much lower than 0.05. It is noteworthy that, if the KS test is performed considering only the samples formed by (ir)relevant subsets **R** of the same size, the hypothesis is rejected with greater confidence for subsets of greater size.

Repeating this experiment using a larger set **H** rather than a singleton $H$ (see Fig. 7), we find that $f(JSD \mid Rel = False)$ is much more uniform, meaning that a high value of maximum JSD no longer necessarily implies relevance (i.e. now we also have irrelevant subsets with a high maximum value of JSD). However, since $f(JSD \mid Rel = True)$ does not change that much with the size of **H**, relevant sets are still more likely to have a higher JSD and, therefore, a low maximum value of JSD is still a sign of irrelevance, even if the contrary is not true (irrelevance does *not* imply a low maximum JSD in this scenario). The null hypothesis of the KS test is still rejected, although with less confidence.

#### 4.4.4. Hill climbing heuristic algorithm

The idea for our heuristic is thus to add to an empty set **R** the node that minimises the maximum value of JSD iteratively, incrementing the size of said set by one in each iteration. When adding any node results in **R** being rendered relevant, we finish the process. This heuristic would reduce the size of the search space from the exponential order that characterised the exhaustive approach to just $O(|\mathbf{S}| \log(|\mathbf{S}|))$. We further assume that the node that is going to minimise the JSD is actually the one with a lower value for the JSD itself. Although this has not been proven in our paper, preliminary experiments show that the results obtained by making this assumption do not vary. As a result, we just need to compute the maximum value of JSD of all the singletons of **S** and iteratively add the ones with the lowest maximum JSD, checking in each step that the resulting subset is still irrelevant. The size of the search space then drops to $O(|\mathbf{S}|)$ and a formalisation can be found in Algorithm 4.

The pruning step that avoids computing sets that contain a relevant subset is implicitly included by first considering the set of irrelevant singletons only and then by not trying to add a node that yields a relevant node set twice. It is also noteworthy that, in the actual implementation, lines 1 and 2 should be coded in parallel in order to compute both (ir)relevance and the JSD using a single evidence propagation.

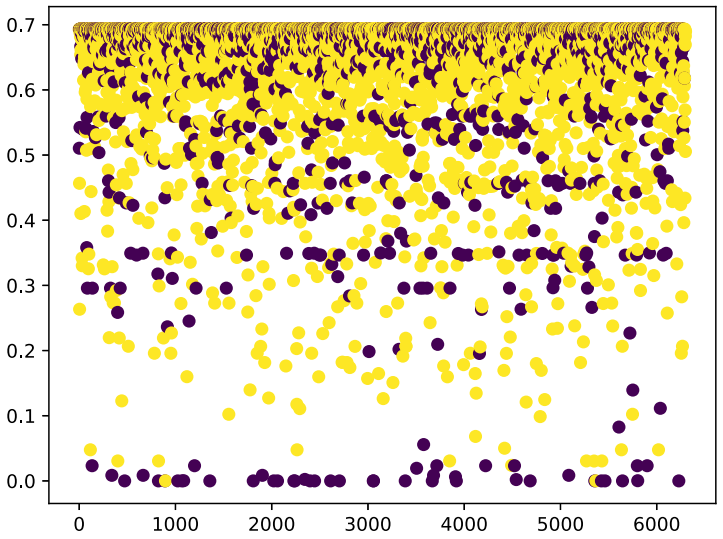


**Fig. 7.** Scatter plot, representing the same experiment as Fig. 6, but with a set **H** of size 5.

**Algorithm 4:** Heuristically inducing the largest irrelevance set for $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$.

**Input :** Bayesian network $\mathcal{B}$ with nodes $\mathbf{V} = \mathbf{H} \cup \mathbf{E} \cup \mathbf{S}$, $\mathbf{h}^* = \text{MAP}(\mathbf{H}, \mathbf{e})$
**Output:** Set $\mathbf{O}$ with the largest irrelevant set

$\mathcal{R} \leftarrow$ compute relevant singletons (Algorithm 1); $\mathcal{I} \leftarrow \mathbf{S} \setminus \mathcal{R}$
Compute $\max_{\mathbf{i} \in \mathbf{I}} \text{JSD}(\Pr(\mathbf{H} \mid \mathbf{e}) \,||\, \Pr(\mathbf{H} \mid \mathbf{i}\,\mathbf{e}))$, $\forall\, \mathbf{I} \in \mathcal{I}$
Sort $\mathcal{I}$ from lowest to highest max. value of JSD
$\mathbf{O} \leftarrow \emptyset$
**forall** $\mathbf{I}_i \in \mathcal{I}$ **do**
  **if** MAP-ind$(\mathbf{O} \cup \{\mathbf{I}_i\}, \mathbf{h}^*, \mathbf{e})$ **then**
    $\mathbf{O} \leftarrow \mathbf{O} \cup \{\mathbf{I}_i\}$
    prune singletons from $\mathcal{I}$ that are d-separated from **H** by **O** and add to **O**
  **end**
**end**
**return O**

An alternative heuristic, based on Proposition 12, would be to try to first add the nodes that are in the Markov blanket of **H**. This would allow us to test closer nodes that can potentially d-separate more nodes from **H** than the average. Each time we add a node to the current best solution **O**, then we will also add all the nodes that are d-separated from **H** given **O** to the set **O**.

## 5. Using relevance in explanations

Although implementation in a more user-centric platform is yet to be established, in this section we provide some theoretical guidelines to study how our algorithms can actually be used for defeasible reasoning-based explanations that convey the stability of a MAP-explanation to an end user. By linking MAP-independence to defeasible reasoning, existing concepts from defeasible reasoning can be employed in the context of Bayesian networks and their explanations. In fact, MAP-independence provides a new framework for defeasible reasoning in Bayesian networks, where we take **e** to be the initial premise, $\mathbf{h}^*$ the conclusion reached and a subset of supplementary nodes **R** to be a defeater if $\mathbf{h}^*$ is not MAP-independent from **R** (i.e., **R** invalidates our conclusion).

We can explain to the user that the MAP-explanation is expected to remain the same as long as relevant nodes, or defeaters, are not observed. Based upon user preferences, we can restrict the explanation to relevant singletons or also return larger sets, e.g., those in $\mathcal{P}^{m+}$, or a selection thereof. If we define a measure of *degree of relevance*, the relevance sets can be ordered from most relevant to least relevant for the stability of the MAP-explanation, and we can select which relevance sets to include in an explanation based on that.

The degree of relevance can be defined in various ways. When discussing the use of FSS filter approaches in Section 4.4 we already mentioned a MAP-independence *strength* measure, which can be defined as the expected (ir)relevance of a set **R** to a MAP-explanation:

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

where $\text{REL}(\mathbf{R}, \mathbf{h}^*, \mathbf{e}) = \{\mathbf{r} \in \Omega(\mathbf{R}) \mid \arg\max_{\mathbf{h} \in \Omega(\mathbf{H})} \Pr(\mathbf{h}\,\mathbf{r} \mid \mathbf{e}) = \mathbf{h}' \text{ for any } \mathbf{h}' \neq \mathbf{h}^*\}$. This measure captures how likely it is that the MAP-explanation will change due to future observations for $\mathbf{R}$; higher values indicate that a set is more relevant and hence less MAP-independent. This strength measure is similar in concept to the *Same-Decision Probability* (SDP) [6].

Another option of defining a degree of relevance is to consider the average probability mass assigned to an alternative MAP-explanation $\mathbf{h}'$ in the case of future observations for $\mathbf{R}$:

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

The suitability of these and similar measures, both in terms of the computational overhead and their ability to distinguish relevance sets, still needs to be investigated.

With some additional bookkeeping, we cannot only indicate which nodes are relevant to the MAP-explanation $\mathbf{h}^*$, and possibly their degree of relevance, but also communicate the specific values $\mathbf{r}$ of the nodes that are relevant to $\mathbf{h}^*$ and the value $\mathbf{h}'$ into which the most likely hypothesis changes upon their observation. This allows us to provide contrastive robustness explanations and detail the values of defeaters:

- any $\mathbf{r}$ found provides for a concrete defeater for $\mathbf{h}^*$;
- any $\mathbf{r}$ found that results in $\mathbf{h}'$ as most likely value, provides an explanation that *contrasts* outputs $\mathbf{h}^*$ and $\mathbf{h}'$.

We would like to note that, similarly to counterfactual explanations, contrastive explanations usually focus on changes in the observed value of the evidence nodes, rather than on effects of possible future observations.

The above discussion shows that 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 criteria that are important for XAI, i.e. good explanations are contrastive, selected, do not refer to probabilities and are social [24]. 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 prioritise the explanations based upon degree of relevance. By interacting with the user, taking into account their preference with respect to the 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 and the defeaters that cause these changes rather than focusing on changes in probabilities.

## 6. Experiments

In this section, we describe our experiments with Algorithms 1, 2 and 4, as well as a use case in which MAP-independence is applied from the defeasible reasoning viewpoint.

### 6.1. Platform for the experiments

For the sake of comprehensibility, we provide the characteristics of the machine used to launch the experiments, as in many of them we measure the execution time. We use

- Ubuntu 22.04.2 LTS as operating system,
- an Intel(R) Core(TM) i7-4790K CPU @ 4.00 GHz processor,
- 16 gigabytes of RAM memory
- and a Radeon RX 560D GPU.

The experiments can be replicated using the notebook available in the GitHub repository of this paper.[4]

### 6.2. Networks used

For the experiments, we use three different networks, motivated by various reasons.

- First, we selected the Insurance network [5], as it is well-known in the field of probabilistic graphical models. The network models the probability of a car accident, involving many variables such as the risk aversion of the driver, the cost of the car, the age, etc. It will be useful to illustrate many of our examples comprehensible and its medium size will serve us to empirically validate many of our hypotheses. The network structure can be visualised in Fig. 5.

[4] https://github.com/Enrique-Val/defeater_explanations.

**Table 1**
Networks used in the experiments. The number of arcs of the Hepatitis network may vary if another learning algorithm is used.

| | Nodes | Arcs |
|---|---|---|
| Insurance | 27 | 52 |
| Hepatitis | 19 | 44 |
| Andes | 223 | 338 |

**Table 2**
Comparison of execution time (mean time and standard deviation in seconds) of the brute force algorithm (BF) and Algorithm 1 (A1) for computing relevant singletons, for different sizes of sets **H** and **S**.

| | Alg | $\lvert\mathbf{S}\rvert = 50$ | $\lvert\mathbf{S}\rvert = 100$ | $\lvert\mathbf{S}\rvert = 150$ | $\lvert\mathbf{S}\rvert = 200$ |
|---|---|---|---|---|---|
| $\lvert\mathbf{H}\rvert = 1$ | BF | $0.22 \pm 0.12$ | $1.10 \pm 0.58$ | $2.40 \pm 0.85$ | $3.20 \pm 0.77$ |
| | A1 | $0.06 \pm 0.02$ | $0.11 \pm 0.02$ | $0.17 \pm 0.04$ | $0.34 \pm 0.04$ |
| $\lvert\mathbf{H}\rvert = 5$ | BF | $0.41 \pm 0.18$ | $1.39 \pm 0.47$ | $2.73 \pm 0.30$ | $3.51 \pm 0.55$ |
| | A1 | $0.37 \pm 0.04$ | $0.66 \pm 0.11$ | $1.24 \pm 0.21$ | $2.78 \pm 0.73$ |
| $\lvert\mathbf{H}\rvert = 10$ | BF | $0.7 \pm 0.1$ | $1.77 \pm 0.25$ | $2.81 \pm 0.27$ | $4.87 \pm 1.79$ |
| | A1 | $11.74 \pm 0.63$ | $20.10 \pm 1.82$ | $34.35 \pm 7.57$ | $86.56 \pm 25.18$ |

- Then we decided to learn a network from scratch using the dataset Hepatitis, that describes whether a hepatitis patient is at risk of dying or not. With this dataset, we introduce a new problem that we find specially interesting. First, the dataset distinguishes between a set of variables for which observations are easier to obtain (namely, physical evidence that can be spotted by a doctor with a brief check) and some other more difficult to observe variables (results of blood tests and prothrombine time tests). If we consider the latter as our set **S**, we would be answering the question "Can I diagnose the severity of the illness by just observing the physical evidence observed in a patient?" We believe that such question will illustrate very concisely the usefulness of MAP-independence and the importance of working on its improvement. Finally, we will be exploring MAP-independence in a more data-driven environment, as the network is not provided, but learned. Specifically, we will use the standard hill climbing algorithm with the K2 score [8] to learn a Bayesian network from this dataset; the dataset was pre-processed by imputing missing values and discretising continuous attributes into bins.
- We also decided to work with the network Andes [7], which acts as an intelligent tutoring system. It is much larger in size, allowing us to get a better insight into simple experiments (for instance, just computing singletons).

The characteristics of all networks are presented in Table 1. Bayesian network computations were done using both GeNIe[5] and the open source library PyAgrum [12]; the Insurance and Andes Bayesian networks were taken from the bnlearn repository[6] and the Hepatitis dataset can be found in the UCI repository[7] [11].

### 6.3. Speed-up of computing singletons for the Andes network

In our first experiment, we aimed to compare the speed-up obtained by applying the ideas presented in Section 4 and formalised in Algorithm 1, in contrast with a brute-force approach to computing MAP-independence for every singleton in the set **S** in a large network. To this end we use the Andes network. We vary the sizes of **S** and **H** and randomise which nodes are included in each set. The evidence nodes **E** are the remaining nodes of the network and the value assignments **e** are sampled for each experiment using the Bayesian network. The results are presented in Table 2.

For small sizes of the set **H**, Algorithm 1 is much faster and scales much better when the number of supplementary nodes $|\mathbf{S}|$ grows large. On the contrary, for larger sizes of **H**, the time of finding the singletons via computing MAP-independence in a brute manner remains quite stable, but there is a significant increase in time for Algorithm 1 when we have large sets **S**. This can be explained because the number of times we enter the loop in line 6 is exponential in $|\mathbf{H}|$ and, furthermore, we are propagating more evidence, making inference more costly.

A solution to get the best of both approaches would be, in case of having a significantly large set of hypothesis **H**, to first execute lines 1 to 5 of Algorithm 1 to find some relevant singletons with a single evidence propagation, and then to verify the (ir)relevance of the remaining singletons by computing MAP-independence for each one of them.

It is also noteworthy that, for a minority of instances, there are very subtle differences in the set of relevant singletons obtained using both algorithms, due to Python internal rounding of numbers. While these differences are almost not noticeable when computing singletons and usually do not even appear for small **H**, if we keep looking for (ir)relevant sets of

[5] www.bayesfusion.com/genie/.
[6] www.bnlearn.com/bnrepository/.
[7] uci.edu/ml/datasets/Hepatitis/.

**Table 3**
For each explanation is shown: the number of nodes d-separated from **H** given **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.

| 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 |
| $\langle \mathbf{h}_{1,2}^*, \mathbf{e}\rangle$ | 2 | 1 | 12 | 7 |

larger sizes, the initially small differences may increase during the run time of the algorithm. Since execution time does not significantly differ between the brute-force approach and Algorithm 1 for small sizes of **H** and **S**, we will compute singletons using the brute force approach in the following experiments to keep coherence in the results, as the networks used are much smaller (Insurance and Hepatitis).

As a final note, we can see that unless both sets **H** and **S** grow extremely large, run times for both algorithms are reasonable. The main issue is to reduce the run time upon trying to find larger relevance sets, since the search space becomes exponential with **S** and checking MAP-independence is much more expensive.

### 6.4. Computing singletons for the Insurance network

We applied the procedure described in Algorithm 1 on the Insurance network shown in Fig. 5. Moreover, we established which nodes fit the different existing notions of irrelevance reviewed in Section 2.2, i.e. d-separated, barren, and nuisance nodes.

The Insurance network contains $|\mathbf{V}| = 27$ nodes, with 3.3 states on average. The graph is multiply connected and the average in-degree of the nodes is 1.9. A total of 1419 probability parameters are specified. We entered evidence **e** for $|\mathbf{E}| = 5$ nodes: Age = `senior`, SocioEcon = `uppermiddle`, DrivHist = `many`, HomeBase = `city`, MakeModel = `sportscar`. We consider three different sets of hypothesis nodes: Accident ($H_1$), whose domain is $\Omega(H_1) = \{$`None`, `Mild`, `Moderate`, `Severe`$\}$; RiskAversion ($H_2$), with $\Omega(H_2) = \{$`Psychopath`, `Adventurous`, `Normal`, `Cautious`$\}$, 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 $\mathbf{h}_{1,2}^*$ for $\{H_1, H_2\}$.

In Table 3 we list, for each explanation context, the number of nodes d-separated from **H** given **E**, the number of nuisance nodes, barren nodes and relevant nodes. For context $\langle h_1^*, \mathbf{e}\rangle$ all 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 $\mathbf{h}_{1,2}^*$ happens to be the union of those relevant to $h_1^*$ and to $h_2^*$.

- For $H_1$ (Accident) we find the same six 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. The six individually relevant nodes are $\mathcal{R}_1 = \{$IliCost, MedCost, OtherCarCost, PropCost, ThisCarCost, ThisCarDam$\}$. Moreover, we also know that $h_1^*$ is weakly MAP-independent of all subsets of the remaining 15 nodes that form $\mathbf{V} \setminus (\mathbf{E} \cup \{H_1\} \cup \mathcal{R}_1)$.
- For $H_2$ (RiskAversion) we find the relevant node $\mathcal{R}_2 =$ SeniorTrain for two out of the three values other than $h_2^*$. As a result, we have that $h_2^*$ is weakly MAP-independent of all subsets of the 20 nodes that form $\mathbf{V} \setminus (\mathbf{E} \cup \{H_2\} \cup \mathcal{R}_2)$.
- For $\{H_1, H_2\}$ we evaluate $4 \cdot 4 - 1 = 15$ configurations. In many of these configurations, we find many relevant sets that are subsets of the one found for $h_1^*$. For instance, in six of these configurations, we find the same six relevant nodes that we found for $h_1^*$; in two cases, we find a subset of five of these six nodes, and for one configuration we find five relevant nodes, four of which are a subset of those found for $h_1^*$ and the fifth being the relevant node for $h_2^*$. Only for one configuration did we find no relevant nodes.

We conclude that $\mathcal{R}_{1,2} = \mathcal{R}_1 \cup \mathcal{R}_2$ is the set of nodes individually relevant to $\mathbf{h}_{1,2}^*$, and that $\mathbf{h}_{1,2}^*$ is weakly MAP-independent of all subsets of $\mathbf{V} \setminus (\mathbf{E} \cup \{H_1, H_2\} \cup \mathcal{R}_{1,2})$.

In our above experiment, we observe that the nodes that are relevant to a most-likely combination of hypotheses are 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 shown in the next example.

**Example 5.** Consider the Bayesian network in Fig. 2, with the following probabilities specified for nodes $A$, $B$ and $C$:

$$\Pr(a) = 0.15 \quad \Pr(c \mid a\,b) = 0.4 \quad \Pr(c \mid a\,\overline{b}) = 0.7$$
$$\Pr(b) = 0.65 \quad \Pr(c \mid \overline{a}\,b) = 0.5 \quad \Pr(c \mid \overline{a}\,\overline{b}) = 0.05$$

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

**Table 4**
Comparison of execution time (mean and standard deviation in seconds) of the brute force algorithm and Algorithm 2 for finding the (ir)relevant sets in $\mathcal{P}(\mathbf{S})$ for $\mathbf{H} = \{\text{Accident}\}$ and different sizes of the set $\mathbf{S}$.

| | $\lvert\mathbf{S}\rvert = 2$ | $\lvert\mathbf{S}\rvert = 4$ | $\lvert\mathbf{S}\rvert = 6$ | $\lvert\mathbf{S}\rvert = 8$ |
|---|---|---|---|---|
| Brute force | 0.01 ± 0.00 | 0.14 ± 0.17 | 1.59 ± 1.57 | 40.44 ± 68.71 |
| Exhaustive (prunings) | 0.00 ± 0.00 | 0.07 ± 0.10 | 0.99 ± 1.69 | 9.17 ± 15.01 |

**Table 5**
Comparison of execution times (mean and standard deviation in seconds) and quality of results of the exhaustive and heuristic algorithms (with the heuristics based on the JSD divergence and on the Markov blanket, MB), with $\mathbf{H} = \{\text{RiskAversion}\}$ and different sizes of the set $\mathbf{S}$. The rows "Optimal ratio" refers to how much larger (in percentage) the largest irrelevant set induced by the exhaustive algorithm is compared to that obtained using the heuristic approach.

| | $\lvert\mathbf{S}\rvert = 2$ | $\lvert\mathbf{S}\rvert = 6$ | $\lvert\mathbf{S}\rvert = 10$ |
|---|---|---|---|
| Exhaustive (pruning) | 0.00 ± 0.00 | 0.92 ± 1.82 | 73.38 ± 111.44 |
| Hill climbing JSD | 0.01 ± 0.01 | 0.75 ± 0.88 | 7.00 ± 5.35 |
| Optimal ratio JSD (%) | 0.00 ± 0.00 | 0.0 ± 0.00 | 5.00 ± 10.00 |
| Hill climbing MB | 0.01 ± 0.01 | 0.47 ± 0.77 | 5.99 ± 6.86 |
| Optimal ratio MB (%) | 0.00 ± 0.00 | 0.00 ± 0.0 | 0.00 ± 0.00 |

### 6.5. Speed-up of the exhaustive and heuristic algorithm

With the experiments in this section we aimed to compare the speed-up obtained by using the exhaustive algorithm with the proposed prunings (see Algorithm 2) and the approximate algorithm. For the approximate algorithm, we used two different heuristics: (1) incrementally add irrelevant singletons with a lower JSD (Hill climbing JSD) and (2) prioritise nodes that are in the Markov blanket of $H \in \mathbf{H}$ (Hill climbing MB); see Section 4.4 and Algorithm 4 for details.

The experiments are performed on the Insurance network, selecting $\mathbf{H} = \{\text{Accident}\}$ and using various sizes of $\mathbf{S}$. Like in Section 6.3, for each experiment, the nodes belonging to $\mathbf{S}$ are randomised and the set $\mathbf{E}$ is formed from the remaining nodes of the network, $\mathbf{E} = \mathbf{V} \setminus \mathbf{H} \cup \mathbf{S}$. Likewise, the value assignment $\mathbf{e}$ is randomised by sampling it using the Bayesian network.

We divide the comparison into two parts. First, the comparison between the brute-force approach (computing MAP-independence for every $\mathbf{R}$ in $\mathcal{P}(\mathbf{S})$) and the exhaustive approaches improved with pruning is summarised in Table 4. We can conclude that pruning in relation to MAP-independence is just plain better and significantly decreases the run time, especially when $|\mathbf{S}|$ increases and computing MAP-independence becomes more and more costly. Note that the sets induced are exactly the same, as both algorithms explore the entire search space.

Second, we compare the exhaustive approach and the approximate algorithm (with both heuristics), leaving out the brute-force approach, as it is clearly outperformed by the exhaustive algorithms with pruning. The results can be seen in Table 5.

The results obtained in these experiments give us a lot of insight about MAP-independence, the potential of exploiting its properties and its computational complexity. On average the heuristic algorithms are faster than the exhaustive counterpart and, in most cases, they find the best solution, i.e., an irrelevant set of the same size as the largest one found by the exhaustive algorithm. It should also be noted that the heuristic based on prioritising sets included in the Markov blanket of $\mathbf{H}$ slightly outperformed the one based on prioritising sets with a low maximum value for JSD. This might be because, in our experiments, the set $\mathbf{H}$ is, in fact, a singleton, and thus there are more chances for a node $V$ to be d-separated from $\mathbf{H}$.

A very insightful phenomenon is the fact that, in many experiments, the exhaustive algorithm finishes faster than both approximate algorithms. This can be explained by the fact that computing MAP-independence for small sets $\mathbf{R}$ is much faster than for large ones. As such, in some cases, the exhaustive algorithm verifies MAP-independence of many small sets $\mathbf{R}$ that are completely ignored by the heuristic algorithm due to its greedy nature, which at times results in large pruning of the search space, thus avoiding to check large sets $\mathbf{R}$. In a paradoxical manner, the approximate algorithm ends up evaluating larger sets $\mathbf{R}$ due to applying less overall pruning. However, it should be noted that, on average, the approximate algorithms still run faster.

### 6.6. An example of defeasible reasoning in the medical domain with the Hepatitis dataset

In this section, we describe an experiment that aims to illustrate the use of MAP-independence for defeasible reasoning. In addition, it serves to illustrate that the concept of MAP-independence can be put to use in the context of test selection [2] To this end, we use the (learned) Hepatitis network.

We consider the scenario in which we are in a hospital evaluating the severeness of the situation of two patients. Each patient will cover a different case: one is (allegedly) a false positive and one a false negative. The characteristics as well as the classification of each patient are described in Table 6.

**Table 6**
Evidence about two hypothetical patients in a hospital.

| | age | sex | steroid | antivirals | fatigue | malaise | anorexia | liver_big |
|---|---|---|---|---|---|---|---|---|
| Pat. 1 | (40, 60] | female | True | False | True | False | False | True |
| Pat. 2 | (40, 60] | female | True | False | False | False | False | False |
| | liver_firm | spleen_palpable | | spiders | ascites | varices | histology | class |
| Pat. 1 | False | True | | True | True | False | True | High risk |
| Pat. 2 | True | False | | True | False | False | True | Low risk |

The table reflects the evidence obtained so far, but there is a set of available tests that may alter the current classification of each patient, namely many levels of substances present in the blood (albumin, bilirubin, SGOT levels, etc.) and a prothrombine time test. For the sake of simplicity, we assume that every test related to blood substances is performed independently. We will use the exhaustive algorithm to determine which tests may yield a result that will invalidate the classification of said patient.

For patient 1 (with a high risk class value), all of the tests fall into the irrelevant category, meaning that they are not necessary to confirm that the patient is at a high risk of dying. No result may change the fact that said patient should be labelled as a high risk patient, confirming that we do not have a false positive. For patient 2 (with lower risk), it is advisable to do some additional tests. Specifically, examining bilirubin levels alongside performing a prothrombine time test would be enough to determine whether we have a true or false negative, meaning that no test related to SGOT levels, alkaline phosphate or albumin needs to be done.

Now suppose that we had no clear evidence about the liver and the steroid use of the patient (and they are considered by all means unobservable). Should we still discard the SGOT levels, alkaline phosphate or albumin tests? The answer is no, as can be checked using MAP-independence. This highlights the importance of having clear evidence in such scenarios.

## 7. Conclusions and further research

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 suggest that a partition of the exponentially large powerset of supplementary nodes is required to find all (ir)relevance sets.

We have first proposed an algorithm for finding relevant singletons that exploits the advantages of having a Bayesian network and then used some of the properties formulated to design an exhaustive algorithm that is able to prune part of the search space. In addition, we have proposed some heuristic alternatives, either analysing a dataset using filters that yield fast but incomplete solutions or using a hill climbing approach, which speed up the computation time significantly in comparison to the exhaustive algorithm at the risk of finding suboptimal solutions. We showed some experiments using the latter and contrasted the result with the exhaustive approach.

In future works, we would like to investigate and detail more efficient approaches to finding (ir)relevance sets, starting with implementing more sophisticated pruning and heuristics. We believe that more complex heuristics such as genetic algorithms should be avoided due to the fact that computing MAP-independence for small subsets of supplementary nodes is much faster and, as such, we suggest to stick to bottom-up approaches. A promising idea would be to apply a beam search, which can be done by launching multiple hill climbers in parallel and pruning cooperatively. We would keep the bottom-up approach and it is still a heuristic algorithm that does not visit the whole search space, but it checks more solutions than hill climbing and, as such, much larger subsets of supplementary nodes will be pruned, getting benefits from both approaches presented in this work. Even though the solutions are optimal even with such a simple heuristic, this might help to reduce the run time. Moreover, we will study and experiment with the proposed measures of degree of relevance in more detail, which have been relegated to a more theoretical study in this paper, as well as further establish the link between defeasible reasoning and MAP-independence. Finally, we would like to evaluate the suggested explanations with actual users.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The paper includes a link to a Github repository with the code. Data and networks used are publicly available and referenced in the paper.

## Acknowledgements

This research was supported by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research[8] (grant number 024.004.022); the Spanish Ministry of Science and Innovation through the PID2019-109247GB-I00 project and the TED2021-131310B-I00 "Bayesian Networks for Interpretable Machine Learning and Optimization (BAYES-INTERPRET)" project, and Ministry of Education through the University Professor Training (FPU) program fellowship,[9] with grant reference FPU21/04812.

## References

[1] E. Albini, A. Rago, P. Baroni, F. Toni, Influence-driven explanations for Bayesian network classifiers, in: Proceedings of the Eighteenth Pacific Rim International Conference on Artificial Intelligence, in: LNAI, vol. 13031, Springer, 2021, pp. 88–100.
[2] S. Andreassen, Planning of therapy and tests in causal probabilistic networks, Artif. Intell. Med. 4 (1992) 227–241.
[3] M. Baker, T.E. Boult, Pruning Bayesian Networks for Efficient Computation, Uncertainty in Artificial Intelligence, vol. 6, Elsevier Science, 1991, pp. 225–232.
[4] C. Bielza, P. Larrañaga, Discrete Bayesian network classifiers: a survey, ACM Comput. Surv. 47 (1) (2014) 1–43.
[5] J. Binder, D. Koller, S. Russell, K. Kanazawa, Adaptive probabilistic networks with hidden variables, Mach. Learn. 29 (1997) 213–244.
[6] A. Choi, Y. Xue, A. Darwiche, Same-decision probability: a confidence measure for threshold-based decisions, Int. J. Approx. Reason. (2012) 1415–1428.
[7] C. Conati, A.S. Gertner, K. VanLehn, M.J. Druzdzel, On-line student modeling for coached problem solving using Bayesian networks, in: Proceedings of the Sixth International Conference on User Modeling, Springer, 1997, pp. 231–242.
[8] G.F. Cooper, E. Herskovits, A Bayesian method for the induction of probabilistic networks from data, Mach. Learn. 9 (1992) 309–347.
[9] I.P. Derks, A. De Waal, A taxonomy of explainable Bayesian networks, in: Proceedings of the Southern African Conference for Artificial Intelligence Research, Springer, 2021, pp. 220–235.
[10] M.J. Druzdzel, H. Suermondt, Relevance in probabilistic models: backyards in a small world, in: Working Notes of the AAAI 1994 Fall Symposium Series: Relevance, 1994, pp. 60–63.
[11] D. Dua, C. Graff, UCI machine learning repository, http://archive.ics.uci.edu/ml, 2017.
[12] G. Ducamp, C. Gonzales, P.-H. Wuillemin, aGrUM/pyAgrum: a toolbox to build models and algorithms for probabilistic graphical models in Python, in: Proceedings of the Tenth International Conference on Probabilistic Graphical Models, in: Proceedings of Machine Learning Research, vol. 138, 2020, pp. 609–612.
[13] J.Y. Halpern, J. Pearl, Causes and explanations: a structural-model approach. Part II: explanations, Br. J. Philos. Sci. 56 (4) (2005) 889–911.
[14] F.V. Jensen, T.D. Nielsen, Bayesian Networks and Decision Graphs, 2nd edition, Springer Science & Business Media, 2007.
[15] R. Koons, Defeasible reasoning, in: The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, 2022.
[16] T. Koopman, S. Renooij, Persuasive contrastive explanations for Bayesian networks, in: Proceedings of the Sixteenth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, in: LNAI, vol. 12897, Springer, 2021, pp. 229–242.
[17] J. Kwisthout, Explainable AI using MAP-independence, in: Proceedings of the Sixteenth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, in: LNAI, vol. 12897, Springer, 2021, pp. 243–254.
[18] J. Kwisthout, Motivating explanations in Bayesian networks using MAP-independence, Int. J. Approx. Reason. 153 (2023) 18–28.
[19] J. Kwisthout, C.P. De Campos, Computational complexity of Bayesian networks, in: Lecture Notes of Tutorials Associated with the Thirty-First Conference on Uncertainty in Artificial Intelligence, 2015.
[20] C. Lacave, F.J. Díez, A review of explanation methods for Bayesian networks, Knowl. Eng. Rev. 17 (2) (2002) 107–127.
[21] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inf. Theory 37 (1) (1991) 145–151.
[22] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, The Springer International Series in Engineering and Computer Science, vol. 454, Springer, 2012.
[23] F.J. Massey Jr., The Kolmogorov-Smirnov test for goodness of fit, J. Am. Stat. Assoc. 46 (253) (1951) 68–78.
[24] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38.
[25] J. Pearl, Probabilistic Reasoning in Intelligent Systems. Networks of Plausible Inference, Morgan Kaufmann, 1988.
[26] J. Pearl, A. Paz, GRAPHOIDS: a graph-based logic for reasoning about relevance relations, Technical Report R-53-L, UCLA Computer Science Department, 1987.
[27] H. Prakken, S. Renooij, Reconstructing causal reasoning about evidence, in: Legal Knowledge and Information Systems, 2001, pp. 131–137.
[28] S. Renooij, Relevance for robust Bayesian network MAP-explanations, in: Proceedings of the Eleventh International Conference on Probabilistic Graphical Models, in: Proceedings of Machine Learning Research, vol. 186, 2022, pp. 13–24.
[29] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215.
[30] Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
[31] R.D. Shachter, Probabilistic inference and influence diagrams, Oper. Res. 36 (1988) 589–604.
[32] I. Stepin, J.M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001.
[33] H.J. Suermondt, Explanation in Bayesian Belief Networks, Phd thesis, Stanford University, 1992.
[34] S.T. Timmer, J.-J.C. Meyer, H. Prakken, S. Renooij, B. Verheij, Explaining Bayesian networks using argumentation, in: Proceedings of the Thirteenth Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer, 2015, pp. 83–92.
[35] S.T. Timmer, J.-J.C. Meyer, H. Prakken, S. Renooij, B. Verheij, A two-phase method for extracting explanatory arguments from Bayesian networks, Int. J. Approx. Reason. 80 (2017) 475–494.
[36] E. Valero-Leal, P. Larrañaga, C. Bielza, Extending map-independence for Bayesian network explainability, in: Proceedings of the Workshop Heterodox Methods for Interpretable and Efficient Artificial Intelligence, Zenodo, 2022, https://doi.org/10.5281/zenodo.7738830.
[37] E. Valero-Leal, P. Larrañaga, C. Bielza, Interpreting time-varying dynamic Bayesian networks for Earth climate modelling, in: Proceedings of the Eleventh International Conference on Probabilistic Graphical Models, in: Proceedings of Machine Learning Research, vol. 186, 2022, pp. 373–384.

[8] https://hybrid-intelligence-centre.nl.
[9] https://www.educacionyfp.gob.es/servicios-al-ciudadano/catalogo/general/99/998758/ficha/998758-informacion-comun.html.