Reintroducing Credal Networks under Epistemic Irrelevance

Jasper De Bock JASPER.DEBOCK@UGENT.BE
Data Science Lab, Ghent University
Ghent (Belgium)

Abstract

A credal network under epistemic irrelevance is a generalised version of a Bayesian network that loosens its two main building blocks. On the one hand, the local probabilities do not have to be specified exactly. On the other hand, the assumptions of independence do not have to hold exactly. Conceptually, these credal networks are elegant and useful. However, in practice, they have long remained very hard to work with, both theoretically and computationally. This paper provides a general introduction to this type of credal networks and presents some promising new theoretical developments that were recently proved using sets of desirable gambles and lower previsions. We explain these developments in terms of probabilities and expectations, thereby making them more easily accessible to the Bayesian network community.

Keywords: credal networks; epistemic irrelevance; irrelevant natural extension; sets of probabilities; lower expectation.

1. Introduction

Bayesian networks (Pearl, 1988) owe their succes to the main feature that all probabilistic graphical models have in common: they are able to model the uncertainty that is associated with large multivariate problems in a manageable way, by combining local uncertainty models with intuitive graph-based independence assumptions. For a Bayesian network, the independence assumptions are derived from a directed acyclic graph and the local uncertainty models are probability distributions.

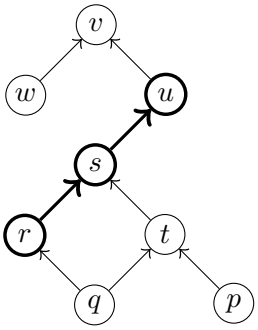
Credal networks (Cozman, 2000, 2005; Antonucci et al., 2014) generalize this concept by replacing the local probability distributions with closed convex sets of probability distributions, also called credal sets. In this way, they do not require the exact specification of all the local probabilities, but allow the user to provide partial constraints on them, such as intervals or inequalities. Depending on the type of credal network that is being considered, the independence assumptions that are derived from the graph are also generalised, by replacing them with weaker types of independence assessments. This paper focusses on credal networks that adopt epistemic irrelevance as their notion of independence, called credal networks under epistemic irrelevance.

To the best of our knowledge, this particular type of credal network was first introduced by Cozman in 1998, be it under a different name—locally defined Quasi-Bayesian network. Now, almost twenty years later, it is firmly established as one of the two main types of credal networks. However, compared with the other main type, which adopts strong independence as its notion of independence, credal networks under epistemic irrelevance have received relatively little attention.

Initial work on credal networks under epistemic irrelevance adopted the framework of probabilities (Cozman, 1998, 2000; de Campos and Cozman, 2007) and, as such, remained close to the theory of Bayesian networks. In contrast, more recent work uses other, closely related frameworks for modelling uncertainty, such as lower previsions and sets of desirable gambles (de Cooman et al., 2010; Benavoli et al., 2011; De Bock and de Cooman, 2014, 2015). However, unfortunately, these

$G = \{p, q, r, s, t, u, v, w\}$

$\Pi(s) = \{r, t\}$
$C(s) = \{u\}$
$D(s) = \{u, v\}$
$N(s) = \{p, q, w\}$
$\Pi N(s) = \{p, q, r, t, w\}$



$K = \{r, s, u\}$

$\Pi(K) = \{q, t\}$
$C(K) = \{v\}$
$D(K) = \{v\}$
$N(s) = \{p, w\}$
$\Pi N(s) = \{p, q, t, w\}$

Figure 1: A simple illustration of several important concepts in this paper.

frameworks are not very well known within the Bayesian network community, and as a direct result, recent work on credal networks under epistemic irrelevance is rather inaccessible to this community.

The aim of this paper is to (re)introduce credal networks under epistemic irrelevance to the Bayesian network community, and to present the most recent theoretical developments in this field in a way that is easily accessible, using the language of probabilities and expectations; lower previsions and sets of desirable gambles are mentioned only in passing. Proofs are not provided, because these do require extensive use of these other frameworks for modelling uncertainty. For a more detailed, technical exposition, which does include proofs, and which also presents a number of algorithms, the interested reader is referred to (De Bock, 2015).

2. The Basics of Credal Networks

Basically, a credal network is just a special type of multivariate uncertainty model for a finite set of variables $\{X_s\}_{s \in G}$, with G some finite index set. Each of the variables X_s takes values x_s in a finite set $\mathcal{X}_s$ and, for any $S \subseteq G$, we use X_S to denote the vector that consists of the variables $\{X_s\}_{s \in S}$, which takes values x_S in $\mathcal{X}_S := \times_{s \in S} \mathcal{X}_s$.

In a credal network, just like in a Bayesian network, the variables $\{X_s\}_{s \in G}$ are identified with their indices $s \in G$. These indices are then interpreted as the nodes of a directed acyclic graph—see Figure 1 for an example—and the arrows of this graph are taken to represent—some type of—(in)dependencies among the individual variables. Finally, these assessments of independence are combined with local uncertainty models, and in this way, a global uncertainty model for X_G is defined. The main difference with a Bayesian network is that the local and global uncertainty models are now *sets* of probability distributions.

In order to formalize this idea, we need some basic graph-theoretic concepts, which are illustrated in Figure 1. For two nodes s and u in G, if there is a directed edge from s to u, we denote this as $s \to u$ and say that s is a *parent* of u and u is a *child* of s. For any node s, its set of parents is denoted by $\Pi(s)$ and its set of children by $C(s)$. A node s is said to *precede* a node v, denoted by $s \sqsubseteq v$, if it is possible to start from s and follow the edges of the graph along their direction to reach v. If $s \sqsubseteq v$ and $s \neq v$, we say that s *strictly precedes* v and write $s \sqsubset v$. For any node s, we call $D(s) := \{v \in G \colon s \sqsubset v\}$ its set of *descendants* and $N(s) := G \setminus (\Pi(s) \cup \{s\} \cup D(s))$ its set of *non-parent non-descendants*. We also use the

shorthand notation $\Pi N(s) := \Pi(s) \cup N(s) = G \setminus (\{s\} \cup D(s))$ to refer to what we call the *non-descendants* of s. With this terminology in place, we can now formally introduce the two main building blocks of a credal network, which are local uncertainty models and assessments of independence.

The basic premise of a credal network is that it is sometimes unrealistic to provide exact values for the local probabilities $P(x_s \mid x_{\Pi(s)})$ that are required to specify a Bayesian network. Therefore, in those cases, the local uncertainty models of a credal network are taken to be sets of probability distributions. For every variable X_s and every instantiation $x_{\Pi(s)}$ of its parent variables $X_{\Pi(s)}$, the associated local uncertainty model is a set $\mathcal{M}_{s|x_{\Pi(s)}}$ of probability mass functions on $\mathcal{X}_s$, the elements of which are regarded as candidates for some ideal—but unknown—conditional probability mass function $P(X_s \mid x_{\Pi(s)})$, in the sense that

$$P(X_s \mid x_{\Pi(s)}) \in \mathcal{M}_{s|x_{\Pi(s)}} \text{ for all } s \in G \text{ and } x_{\Pi(s)} \in \mathcal{X}_{\Pi(s)}. \tag{1}$$

Most authors require the sets $\mathcal{M}_{s|x_{\Pi(s)}}$ to be closed and convex, and then call them *credal sets*; we will follow this convention here as well. In practice, these credal sets are elicited from experts, learned from data, or constructed as some type of neighbourhood model. We do not discuss these practical aspects here, but assume that the local credal sets are given.

The second building block of a credal network is a collection of independence assessments. As in Bayesian networks, these independence assessments are inferred from the graph of the network in the following way: every variable X_s is assumed to be conditionally independent of its non-parent non-descendants $X_{N(s)}$ given its parents $X_{\Pi(s)}$. However, in the context of sets of probability distributions, there is no consensus on what is meant here by independence. Depending on the notion of independence that is chosen, a different type of credal network is obtained.

3. Credal Networks under Complete Independence

The most straightforward way to define independence for a set of distributions, is to simply impose the usual notion of independence, which we will henceforth call *stochastic independence*, to each of its elements $P(X_G)$. This type of independence—element-wise stochastic independence—is called *complete independence*. In the case of credal networks, this results in the following assessment:

$$P(X_s \mid x_{\Pi N(s)}) = P(X_s \mid x_{\Pi(s)}) \text{ for all } s \in G \text{ and } x_{\Pi N(s)} \in \mathcal{X}_{\Pi N(s)}. \tag{2}$$

Conventionally, the conditional probabilities in this expression are taken to be derived from $P(X_G)$ through Bayes's rule. However, this creates issues in the case of probability zero; for example, if $P(x_{\Pi N(s)}) = 0$, then $P(X_s \mid x_{\Pi N(s)})$ is ill-defined.

In order to avoid these issues in an elegant yet rigorous way, we will not regard conditional probabilities as derived concepts that are obtained through Bayes's rule, but rather as primitive notions that are part of a (full) conditional probability measure (Dubins, 1975).

Definition 1 *A full conditional probability measure P on a finite set Ω is a map*

$$P\colon \mathcal{P}(\Omega) \times \mathcal{P}_\emptyset(\Omega) \to \mathbb{R}\colon (A, B) \to P(A \mid B),$$

with $\mathcal{P}_\emptyset(\Omega) := \mathcal{P}(\Omega) \setminus \{\emptyset\}$, such that for any $A, C \in \mathcal{P}(\Omega)$ and $B \in \mathcal{P}_\emptyset(\Omega)$:

F1: *$P(\,\cdot\,|B)$ is a probability measure on $\mathcal{P}(\Omega)$ with $P(B|B) = 1$;*

F2: $P(A \cap C|B) = P(A|C \cap B)P(C|B)$ *if* $C \cap B \neq \emptyset$.

The axioms F1 and F2 correspond to the usual rules of probability. The only difference is that Bayes's rule—F2—does not define conditional probabilities by means of division, but instead regards them as primitive notions and requires them to satisfy a product rule.

In the case of a credal network, we model the uncertainty about X_G by means of a—possibly partially specified—full conditional probability measure P on $\mathcal{X}_G$. For any event $A \in \mathcal{P}(\mathcal{X}_G)$ and any non-empty event $B \in \mathcal{P}_\emptyset(\mathcal{X}_G)$, $P(A|B)$ is the probability of A conditional on B and $P(A) \coloneqq P(A|\mathcal{X}_G)$ is the (unconditional) probability of A. We will mostly focus on events of the form $\{z_G \in \mathcal{X}_G \colon z_S = x_S\}$, with $S \subseteq G$; for ease of notation, we denote these events as x_S. The events x_s, $x_{\Pi(s)}$, $x_{\Pi N(s)}$, x_G and $x_\emptyset = \mathcal{X}_G$ correspond to special cases.

Within this framework, Equation (2) is now simply a constraint on the full conditional probability measure P and does not require—nor suffers from—divisions by zero. As is very well known from the theory of Bayesian networks, this constraint implies that the unconditional global probabilities $P(x_G)$ are completely determined by the local probabilities $P(x_s \,|\, x_{\Pi(s)})$:

$$P(x_G) = \prod_{s \in G} P(x_s \,|\, x_{\Pi(s)}) \text{ for all } x_G \in \mathcal{X}_G.$$

This is not necessarily true for conditional probabilities. For example, there may be some $S, T \subseteq G$, $x_S \in \mathcal{X}_S$ and $x_T \in \mathcal{X}_T$ such that $P(x_S \,|\, x_T)$ is not uniquely determined by Equation (2) and the local probabilities. However, this is usually ignored; the theory of Bayesian networks focusses on cases where $P(x_T) > 0$, which guarantees that $P(x_S \,|\, x_T)$ can be computed by means of Bayes's rule—F2. This restricted focus is unfortunate because, even if x_T has probability zero, $P(x_S \,|\, x_T)$ is often still uniquely determined by the local probabilities and Equation (2).

In any case, for credal networks, probability zero is not the only source of non-uniqueness. Indeed, the local probabilities may themselves not be unique, because Equation (1) will in general—unless all the local credal sets are singletons—only impose partial constraints on the local probabilities. Due to this inherent non-uniqueness, a credal network does not correspond to a single full conditional probability measure, but rather to a set of them. We will denote such a set of full conditional probability measures on $\mathcal{X}_G$ by $\mathcal{F}_G$, and for any $B \in \mathcal{P}_\emptyset(\mathcal{X}_G)$ and any $S \subseteq G$, we will then use $\mathcal{F}_G(X_S \,|\, B)$ to refer to the set of probability mass functions $\{P(X_S \,|\, B) \colon P \in \mathcal{F}_G\}$. The sets $\mathcal{F}_G(X_s \,|\, x_{\Pi(s)})$, $\mathcal{F}_G(X_s \,|\, x_{\Pi N(s)})$ and $\mathcal{F}_G(X_G) \coloneqq \mathcal{F}_G(X_G \,|\, \mathcal{X}_G)$ correspond to important special cases.

The largest set of (full conditional) probability measures that is compatible with the defining constraints of a credal network is called its *extension*. For a credal network under complete independence, these defining constraints are Equations (1) and (2), and the corresponding extension is called the *complete extension*. We will denote this complete extension by $\mathcal{F}_G^{\text{com}}$. Clearly, if we let $\mathcal{F}_G^*$ be the set of all full conditional probability measures on $\mathcal{X}_G$, then $\mathcal{F}_G^{\text{com}}$ is given by

$$\mathcal{F}_G^{\text{com}} = \Big\{ P \in \mathcal{F}_G^* \colon \big(\forall s \in G\big)\big(\forall x_{\Pi N(s)} \in \mathcal{X}_{\Pi N(s)}\big) \\ P(X_s \,|\, x_{\Pi N(s)}) = P(X_s \,|\, x_{\Pi(s)}) \in \mathcal{M}_{s|x_{\Pi(s)}} \Big\}. \quad (3)$$

If we make abstraction of the 'full conditional' aspects that we have added, and focus on the unconditional part $\mathcal{F}_G^{\text{com}}(X_G)$, then this complete extension is simply the set of all Bayesian networks

whose local probability mass functions $P(X_s \mid x_{\Pi(s)})$ take values in the local credal sets $\mathcal{M}_{s|x_{\Pi(s)}}$. This approach is highly intuitive if one is convinced that the uncertainty about X_G can be modelled by means of a single Bayesian network and, for some reason, the required local probability mass functions are not exactly known, but are only partially specified. It seems reasonable to model this type of situation by means of a set of Bayesian models, and it should therefore not be surprising that credal networks under complete independence were the first type of credal network to be considered; see (Fagiuoli and Zaffalon, 1998). In fact, at that point in time, there were no other types, and credal networks under complete independence were simply called credal networks.

Nevertheless, today, rather surprisingly, credal networks under complete independence are almost never considered. Instead, the majority of work on credal networks considers what are called credal networks under *strong independence*. We will not go into details here; for our present purposes, it suffices to know that the unconditional part $\mathcal{F}_G^{\text{str}}(X_G)$ of the extension of a credal network under strong independence, which is caled the strong extension of the network, is equal to the convex hull of the unconditional part $\mathcal{F}_G^{\text{com}}(X_G)$ of the complete extension. Given the choice between these two extensions, I favour the complete extension, because of its clear and intuitive sensitivity analysis interpretation, which the strong extension does not have. I fail to understand why most authors prefer the strong extension instead. In any case, the choice is mainly a philosophical one, because in practice, there is little difference between the two approaches. Essentially, Fagiuoli and Zaffalon (1998, Theorem 5) already showed that for many commonly considered parameters of interest—such as (conditional) lower and upper probabilities and expectations—it makes no difference whether we compute them with respect to the complete extension or its convex hull—the strong extension. Hence, most of the algorithms that have been developed for strong extensions can be applied to complete extensions as well.

4. Credal Networks under Epistemic Irrelevance

In this paper, we do not impose complete or strong independence, but instead impose the following assessments of epistemic irrelevance:

$$\mathcal{F}_G(X_s \mid x_{\Pi(s)}) = \mathcal{F}_G(X_s \mid x_{\Pi N(s)}) \text{ for all } s \in G \text{ and } x_{\Pi N(s)} \in \mathcal{X}_{\Pi N(s)}. \tag{4}$$

The idea here is that since we are modelling uncertainty by means of sets of (full conditional) probability measures, independence should be a statement about such sets, not about the individual elements of these sets. Equation (4) imposes that given the value $x_{\Pi(s)}$ of its parents, our beliefs about the variable X_s remain identical if we are also given the value $x_{N(s)}$ of its non-parent non-descendants. The only difference with the more conventional notions of—stochastic, complete or strong—independence lies in the fact that beliefs are now no longer identified with individual probability distributions, but rather with the information that is available about these distributions, that is, with sets of probabilities.

By combining our assessments of epistemic irrelevance with the assessments that are imposed by the local credal sets, we obtain a credal network under epistemic irrelevance. The largest set $\mathcal{F}_G$ of full conditional probability measures on $\mathcal{X}_G$ that satisfies these two assessments—Equations (1) and (4)—is called the *irrelevant natural extension* of a credal network. We will denote this extension by $\mathcal{F}_G^{\text{irr}}$. As shown in (Cozman, 2000) under strict positivity conditions, and more generally in (De Bock, 2015), $\mathcal{F}_G^{\text{irr}}$ is given by

$$\mathcal{F}_G^{\text{irr}} = \Big\{ P \in \mathcal{F}_G^* \colon \big(\forall s \in G\big)\, \big(\forall x_{\Pi N(s)} \in \mathcal{X}_{\Pi N(s)}\big)\, P(X_s \mid x_{\Pi N(s)}) \in \mathcal{M}_{s|x_{\Pi(s)}} \Big\}. \tag{5}$$

If we compare this expression with Equation (3), we see that the defining assessments of a credal network under epistemic irrelevance are less stringent than those of a credal network under complete—or strong—independence. Actually, it can be shown that $\mathcal{F}_G^{\text{com}}$ satisfies the epistemic irrelevance assessments in Equation (4). However, the converse is not true: the full conditional probability measures in $\mathcal{F}_G^{\text{irr}}$ do not need to satisfy Equation (2). In other words, a credal network under epistemic irrelevance does not impose stochastic independence: $P(X_s \mid x_{\Pi N(s)})$ may vary as $x_{N(s)}$ changes, as long as it remains within the local credal set $\mathcal{M}_{s|x_{\Pi(s)}}$. In this sense, epistemic irrelevance imposes a notion of 'almost' stochastic independence, which, in those applications where stochastic independence is an approximation that is imposed out of mathematical convenience, can provide a more realistic alternative.

From a practical point of view, the main object of interest is not the irrelevant natural extension $\mathcal{F}_G^{\text{irr}}$ itself, but rather the corresponding bounds on some parameters of interest, such as probabilities and expected values. The most important such bounds are tight lower and upper bounds on conditional expectations $E(f(X_S) \mid B) \coloneqq \sum_{x_S \in \mathcal{X}_S} f(x_S) P(x_S \mid B)$, where f belongs to the set $\mathcal{G}(\mathcal{X}_S)$ of all real-valued functions on $\mathcal{X}_S$, with S a subset of G, and where $B \in \mathcal{P}_\emptyset(\mathcal{X}_G)$ is a non-empty event. The tightest lower bound on this conditional expectation is the *lower expectation* of f, defined by

$$\underline{E}_G^{\text{irr}}(f(X_S) \mid B) \coloneqq \inf\left\{E(f(X_S) \mid B) \colon P \in \mathcal{F}_G^{\text{irr}}\right\}. \tag{6}$$

The *upper expectation* $\overline{E}_G^{\text{irr}}(f(X_G) \mid B)$ can be defined analogously, simply by replacing the infimum with a supremum. However, because $\overline{E}_G^{\text{irr}}(f(X_S) \mid B) = -\underline{E}_G^{\text{irr}}(-f(X_S) \mid B)$, it suffices to focus on lower expectations. Lower and upper probabilities are defined similarly, but these too will not be our main focus, because they correspond to the special case where f is the indicator $\mathbb{I}_A$ of an event $A \in \mathcal{P}(\mathcal{X}_G)$, defined by $\mathbb{I}_A(x_G) \coloneqq 1$ if $x_G \in A$ and $\mathbb{I}_A(x_G) \coloneqq 0$ otherwise. Indeed, since $P(A \mid B)$ is clearly equal to $E(\mathbb{I}_A(X_G) \mid B)$, it follows that the conditional lower probability $\underline{P}_G^{\text{irr}}(A \mid B) \coloneqq \inf\left\{P(A \mid B) \colon P \in \mathcal{F}_G^{\text{irr}}\right\}$ is equal to $\underline{E}_G^{\text{irr}}(\mathbb{I}_A(X_G) \mid B)$, and similarly for the conditional upper probability $\overline{P}_G^{\text{irr}}(A \mid B)$. Unconditional lower and upper expectations and probabilities are obtained by choosing $B = \mathcal{X}_G$, in which case we will drop the conditional event from the notation, for example by writing $\underline{E}_G^{\text{irr}}(f(X_G))$ instead of $\underline{E}_G^{\text{irr}}(f(X_G) \mid \mathcal{X}_G)$.

Since all the other bounds can be easily derived from them, algorithmic efforts can focus on computing $\underline{E}_G^{\text{irr}}(f(X_G) \mid B)$, which, as can be seen from Equations (5) and (6), requires solving a large optimisation problem. If the local models $\mathcal{M}_{s|x_{\Pi(s)}}$ are described by linear constraints, this optimisation problem can be solved by means of linear programming methods (Cozman, 2000; de Campos and Cozman, 2007). However, unfortunately, the size of the required linear programs is exponential in the size of the network, and therefore, this direct approach only works for small networks. As a result, it has long been thought that credal networks under epistemic irrelevance are computationally intractable.

This perception has recently changed, due to some successful algorithmic developments in terms of lower previsions, which are basically just lower expectations, but with a different interpretation attached to them. For credal networks under epistemic irrelevance of which the graph is a tree, there is now a polynomial-time updating algorithm that can compute $\underline{E}_G^{\text{irr}}(f(X_s) \mid x_T)$ for $s \in G$ and $T \subseteq G \backslash \{s\}$ (de Cooman et al., 2010). This is rather remarkable, especially since the same inference problem is NP-hard for credal networks under strong (or complete) independence (Mauá et al., 2014). Other recent algorithmic developments considered the case of imprecise hidden Markov models under epistemic irrelevance, the graph of which is again a—special type of—tree (Benavoli et al., 2011; De Bock and de Cooman, 2014).

5. Decomposition Properties: Beyond the Special Case of Trees

It is no coincidence that all the recent algorithmic successes with credal networks under epistemic irrelevance have been obtained for networks whose graph is a tree. Essentially, all of these algorithms are based on the fact that for trees, the irrelevant natural extension satisfies a number of convenient theoretical properties, which allow for large computational problems to be decomposed into smaller ones and, as such, enable the development of efficient recursive algorithms; see for example (de Cooman et al., 2010).

In order to develop efficient algorithms for networks that are more general than trees, an important first step is therefore to generalise these theoretical properties from trees to arbitrary directed acyclic graphs. Recently, it has been shown that this is indeed possible (De Bock and de Cooman, 2013, 2015; De Bock, 2015). However, these generalised properties have been obtained and stated using sets of desirable gambles, and are therefore rather inaccessible to the general Bayesian network community. In order to remedy this situation, we here present them in a more accessible format, in terms of lower expectations.

The following property is perhaps the most important one, as it has been—and, no doubt, will remain to be—the backbone of recursive algorithms.

Proposition 2 *Consider any set $S \subseteq G$, with $T \coloneqq G \setminus S$, such that $t \sqsubset s$ for all $t \in T$ and $s \in S$. Then*

$$\underline{E}^{\text{irr}}_G(f(X_G)) = \underline{E}^{\text{irr}}_G(\underline{E}^{\text{irr}}_G(f(X_G) \,|\, X_T)) \text{ for all } f \in \mathcal{G}(\mathcal{X}_G).$$

Basically, this result is just an extension of the law of iterated expectation, or equivalently, the law of total probability. For readers that are acquainted with imprecise-probabilistic jargon: this result establishes marginal extension. An illustration of this perhaps rather abstract property can be found in the example at the end of this section; see Equation (7).

In order to state the next set of results, we need some additional graph-theoretic concepts, which are again illustrated in Figure 1. For any subset K of G, we define its set of parents as $\Pi(K) \coloneqq (\bigcup_{s\in K} \Pi(s)) \setminus K$ and let $D(K) \coloneqq (\bigcup_{s\in K} D(s)) \setminus K$ be its set of descendants. The set of non-descendants of K is given by $\Pi N(K) \coloneqq G \setminus (K \cup D(K))$, and we also define the set $N(K) \coloneqq \Pi N(K) \setminus \Pi(K)$. If K is a singleton $\{s\}$, these concepts reduce to the simple versions in Section 2. Finally, we call a subset K of G *closed* if, for all $s, t \in K$ and $k \in G$, $s \sqsubseteq k \sqsubseteq t$ implies that $k \in K$.

The following proposition establishes a first crucial property for these closed sets.

Proposition 3 *Let K be a closed subset of G. Then for any $f \in \mathcal{G}(\mathcal{X}_K)$, $x_{\Pi(K)} \in \mathcal{X}_{\Pi(K)}$, $h \in \mathcal{G}(\mathcal{X}_{\Pi N(K)})$ and any non-negative $g \in \mathcal{G}(\mathcal{X}_{N(K)})$:*

$$\begin{aligned}\underline{E}^{\text{irr}}_G\big(h(X_{\Pi N(K)}) + g(X_{N(K)})\mathbb{I}_{x_{\Pi(K)}}(X_{\Pi(K)})f(X_K)\big) \\ = \underline{E}^{\text{irr}}_G\big(h(X_{\Pi N(K)}) + g(X_{N(K)})\mathbb{I}_{x_{\Pi(K)}}(X_{\Pi(K)})\underline{E}^{\text{irr}}_G(f(X_K) \,|\, x_{\Pi(K)})\big).\end{aligned}$$

At first sight, this result might seem a bit complicated, but upon closer inspection, it should become clear that it is in fact not. The essential feature here is that the left hand side is a lower expectation of a function that depends on all the variables $\{X_s\}_{s\in G}$, whereas the right hand side consists of two separate lower expectations, each of which depends on fewer variables. The following two corollaries of Proposition 3 highlight this feature even more.

Corollary 4 (factorisation) *Let K be a closed subset of G. Then for any $f \in \mathcal{G}(\mathcal{X}_K)$ and $x_{\Pi(K)} \in \mathcal{X}_{\Pi(K)}$ and any non-negative $g \in \mathcal{G}(\mathcal{X}_{N(K)})$:*

$$\underline{E}_G^{\text{irr}}\big(g(X_{N(K)})\mathbb{I}_{x_{\Pi(K)}}(X_{\Pi(K)})f(X_K)\big) = \begin{cases} \underline{E}_G^{\text{irr}}\big(f(X_K) \,|\, x_{\Pi(K)}\big)\, \underline{E}_G^{\text{irr}}\big(g(X_{N(K)})\mathbb{I}_{x_{\Pi(K)}}(X_{\Pi(K)})\big) & \text{if } \underline{E}_G^{\text{irr}}\big(f(X_K) \,|\, x_{\Pi(K)}\big) \geq 0 \\ \underline{E}_G^{\text{irr}}\big(f(X_K) \,|\, x_{\Pi(K)}\big)\, \overline{E}_G^{\text{irr}}\big(g(X_{N(K)})\mathbb{I}_{x_{\Pi(K)}}(X_{\Pi(K)})\big) & \text{if } \underline{E}_G^{\text{irr}}\big(f(X_K) \,|\, x_{\Pi(K)}\big) \leq 0. \end{cases}$$

Corollary 5 (external additivity) *Let K be a closed subset of G such that $\Pi(k) \subseteq K$ for all $k \in K$. Then for any $f \in \mathcal{G}(\mathcal{X}_K)$ and $h \in \mathcal{G}(\mathcal{X}_{N(K)})$:*

$$\underline{E}_G^{\text{irr}}\big(h(X_{N(K)}) + f(X_K)\big) = \underline{E}_G^{\text{irr}}\big(h(X_{N(K)})\big) + \underline{E}_G^{\text{irr}}\big(f(X_K)\big).$$

A second crucial property for closed sets is that the corresponding so-called sub-network is closely related to the original network. However, before we can formally state this result, we first need to explain what we mean by a sub-network.

With any subset K of G and any fixed value $x_{\Pi(K)}$ of $X_{\Pi(K)}$, we associate a new credal network, called *sub-network*. The graph of this sub-network is obtained from the original graph by simply removing the nodes that do not belong to K and the arrows that enter these nodes or depart from these nodes. For example, in Figure 1, the sub-graph that corresponds to $K := \{r, s, u\}$ is highlighted by means of thicker lines. The local credal sets of the sub-network are equal to the original ones. However, this is not immediate: as illustrated in Figure 1, a node $s \in K$ might have a parent t that does not belong to the sub-graph of K. For this reason, in order to obtain local models that only depend on parents that belong to K, we fix the value x_t of X_t for any $t \in G \setminus K$ that has a child in K, or equivalently, we fix the value $x_{\Pi(K)}$ of $X_{\Pi(K)}$.

For any choice of $K \subseteq G$ and $x_{\Pi(K)}$, the corresponding sub-network, like any credal network, has an irrelevant natural extension. We will denote this irrelevant natural extension by $\mathcal{F}^{\text{irr}}_{K|x_{\Pi(K)}}$ and will use $\underline{E}^{\text{irr}}_{K|x_{\Pi(K)}}$ to refer to the corresponding lower expectations. If K is closed, these lower expectations satisfy the following property.

Proposition 6 *Consider any closed subset K of G and any $x_{\Pi(K)} \in \mathcal{X}_{\Pi(K)}$. Then*

$$\underline{E}_G^{\text{irr}}\big(f(X_K) \,|\, B_K, x_{\Pi(K)}, B_{N(K)}\big) = \underline{E}^{\text{irr}}_{K|x_{\Pi(K)}}\big(f(X_K) \,|\, B_K\big) = \underline{E}_G^{\text{irr}}\big(f(X_K) \,|\, B_K, x_{\Pi(K)}\big)$$

for all $f \in \mathcal{G}(\mathcal{X}_K)$, $B_K \in \mathcal{P}_\emptyset(\mathcal{X}_K)$ and $B_{N(K)} \in \mathcal{P}_\emptyset(\mathcal{X}_{N(K)})$.

The crucial feature of this result is that it allows us to reduce an optimisation problem in the original credal network into a similar but smaller-sized optimisation problem in one of its sub-networks. Propositions 2 and 3 and the corollaries of the latter can then again be applied to these sub-networks, and by continuing in this way, it is possible to reduce large optimisation problems to a combination of multiple small ones, thereby allowing for tractable computations. The following example illustrates how this works for a very simple inference problem. More involved examples can be found in (De Bock, 2015).

Example 1 *Suppose that we are interested in computing $\underline{E}_G^{\text{irr}}\big(\alpha(X_r)+\beta(X_w)\big)$ for a credal network whose graph is depicted in Figure 1, with $\alpha(X_r)$ and $\beta(X_w)$ real-valued functions of X_r and X_w,*

respectively. Proposition 6 then allows us to reduce this problem to a similar problem in a much smaller network: since $G' = \{w, r, q\}$ is a closed subset of G, we find that

$$\underline{E}^{\mathrm{irr}}_{G}\big(\alpha(X_r) + \beta(X_w)\big) = \underline{E}^{\mathrm{irr}}_{G'}\big(\alpha(X_r) + \beta(X_w)\big),$$

where $\underline{E}^{\mathrm{irr}}_{G'}$ is the lower expectation operator of a credal network with only three nodes (w, r and q) and a single edge ($q \to r$). Furthermore, because of Corollary 5, we also find that

$$\underline{E}^{\mathrm{irr}}_{G'}\big(\alpha(X_r) + \beta(X_w)\big) = \underline{E}^{\mathrm{irr}}_{G'}\big(\alpha(X_r)\big) + \underline{E}^{\mathrm{irr}}_{G'}\big(\beta(X_w)\big).$$

The two terms in this sum can now be simplified even more, because Proposition 6 implies that

$$\underline{E}^{\mathrm{irr}}_{G'}\big(\alpha(X_r)\big) = \underline{E}^{\mathrm{irr}}_{G''}\big(\alpha(X_r)\big) \quad \textit{and} \quad \underline{E}^{\mathrm{irr}}_{G'}\big(\beta(X_w)\big) = \underline{E}^{\mathrm{irr}}_{w}\big(\beta(X_w)\big),$$

with $G'' \coloneqq \{r, q\}$. Hence, we have managed to reduce our original problem to two smaller problems: computing $\underline{E}^{\mathrm{irr}}_{G''}\big(\alpha(X_r)\big)$ and $\underline{E}^{\mathrm{irr}}_{w}\big(\beta(X_w)\big)$. Computing $\underline{E}^{\mathrm{irr}}_{w}\big(\beta(X_w)\big)$ is trivial, because it corresponds to a network with a single node w. For $\underline{E}^{\mathrm{irr}}_{w}\big(\beta(X_w)\big)$, Proposition 2 implies that

$$\underline{E}^{\mathrm{irr}}_{G''}\big(\alpha(X_r)\big) = \underline{E}^{\mathrm{irr}}_{G''}\big(\underline{E}^{\mathrm{irr}}_{G''}\big(\alpha(X_r) \mid X_q\big)\big) = \underline{E}^{\mathrm{irr}}_{G''}\big(\gamma(X_q)\big), \tag{7}$$

where we let $\gamma(X_q) \coloneqq \underline{E}^{\mathrm{irr}}_{G''}\big(\alpha(X_r) \mid X_q\big)$. The last simplifying step now consists in a final application of Proposition 6, from wich we infer that

$$\underline{E}^{\mathrm{irr}}_{G''}\big(\gamma(X_q)\big) = \underline{E}^{\mathrm{irr}}_{q}\big(\gamma(X_q)\big) \;\; \textit{and} \;\; (\forall x_q \in \mathcal{X}_q) \;\; \gamma(x_q) = \underline{E}^{\mathrm{irr}}_{r|x_q}\big(\psi(X_r)\big).$$

As before, computing $\underline{E}^{\mathrm{irr}}_{q}\big(\gamma(X_q)\big)$ and $\underline{E}^{\mathrm{irr}}_{r|x_q}\big(\psi(X_r)\big)$ is trivial, because each of these problems corresponds to a network with a single node. The original problem has therefore been reduced to several smaller problems, each of which requires only local optimisations.

6. Epistemic h-irrelevance, AD-separation and Graphoid Axioms

As the reader may have noticed, the second equality in Proposition 6 is redundant: it follows from the first equality by choosing $B_{N(K)} = \mathcal{X}_{N(K)}$. The reason why we nevertheless state it explicitly, is because it illustrates that the irrelevant natural extension satisfies many more epistemic irrelevances than the basic ones that were used to define it in Equation (4). In fact, it even satisfies statements of *epistemic h-irrelevance*.

Definition 7 *(Cozman, 2013) For three pairwise disjoint sets $I, S, C \subseteq G$, we say that X_I is (epistemically) h-irrelevant to X_S conditional on X_C, and write* $\mathrm{HIR}(I, S \mid C)$*, if*

$$\underline{E}^{\mathrm{irr}}_{G}\big(f(X_S) \mid B_S, x_C, B_I\big) = \underline{E}^{\mathrm{irr}}_{G}\big(f(X_S) \mid B_S, x_C\big)$$

for all $f \in \mathcal{G}(\mathcal{X}_S)$, $B_S \in \mathcal{P}_\emptyset(\mathcal{X}_S)$, $x_C \in \mathcal{X}_C$ and $B_I \in \mathcal{P}_\emptyset(\mathcal{X}_I)$.

Indeed, Proposition 6 clearly implies that for any closed subset K of G, $X_{N(K)}$ is epistemically h-irrelevant to X_K conditional on $X_{\Pi(K)}$: $\mathrm{HIR}(N(K), K \mid \Pi(K))$.

This statement of epistemic h-irrelevance is similar to the assessment of epistemic irrelevance that was imposed in Equation (4)—for $S = \{s\}$, $C = \Pi(s)$ and $I = N(s)$—but differs on several

levels. First of all: it is stated in terms of lower expectations, whereas Equation (4) was stated in terms of sets of probabilities. However, this is not so important; epistemic h-irrelevance can also be defined in terms of sets of probabilities (De Bock, 2015). What really sets epistemic h-irrelevance apart from epistemic irrelevance is that is far more powerful when it comes to conditional models. Unlike epistemic irrelevance, as can be seen from Definition 7, epistemic h-irrelevance requires *all* information about the value of X_I—including *partial* information—to be irrelevant to *all* beliefs about X_S—conditional *and* unconditional beliefs—conditional on the value of X_C.

As it turns out, the irrelevant natural extension satisfies many more statements of epistemic h-irrelevance than the ones that are implied by Proposition 6. Similarly to how for a Bayesian network, the well-known d-separation criterion implies stochastic independence, for the irrelevant natural extension, the so-called AD-separation criterion (De Bock, 2015; De Bock and de Cooman, 2015) implies epistemic h-irrelevance.

Definition 8 *For any pairwise disjoint sets $I, S, C \subseteq G$, we say that I is AD-separated from S by C, and write $\mathrm{AD}(I, S \mid C)$, if there is some closed subset K of G such that $S \subseteq K$, $\Pi(K) \subseteq C$, $I \subseteq N(K)$ and $D(K) \cap C = \emptyset$.*

Proposition 9 *For any pairwise disjoint sets $I, S, C \subseteq G$ such that $\mathrm{AD}(I, S \mid C)$, the irrelevant natural extension $\underline{E}_G^{\mathrm{irr}}$ satisfies $\mathrm{HIR}(I, S \mid C)$.*

As a simple illustration of AD-separation, consider for example the sets $I = \{p, w\}$, $S = \{r, u\}$ and $C = \{s, q, t\}$. Then for the DAG in Figure 1, by applying Definition 8 for $K = \{r, s, u\}$, we find that I is AD-separated from S by C.

The Bayesian network counterpart of Proposition 9—with AD-separation and epistemic h-irrelevance replaced by d-separation and stochastic independence—is proved by exploiting the fact that stochastic independence satisfies various graphoid properties (symmetry, redundancy, decomposition,weak union, contraction and intersection). Therefore, since epistemic irrelevance fails some of these graphoid properties (Cozman and Walley, 2005), it has long been thought that a similar result would not hold for credal networks under epistemic irrelevance. However, as Proposition 9 shows, it is nevertheless possible to prove such a result. In order to do so, two steps were essential. The first step was to drop the symmetry of the separation criterion; since epistemic irrelevance is asymmetric, symmetry is not to be expected anyway. The 'A' in AD-separation is short for 'asymmetric', and it can be shown that Definition 8 is indeed a proper asymmetrical version of d-separation (De Bock, 2015). The second step was to not focus on graphoid properties, but use other means to prove separation; for Propositions 6 and 9—the proof of the latter is heavily based on the former—these means were sets of desirable gambles (De Bock and de Cooman, 2015).

The fact that Proposition 9 can be proven without the use of graphoid properties illustrates nicely that these properties are not essential, and that the fact that a notion of independence—such as epistemic irrelevance—fails some of them, should not be regarded as problematic. In fact, I think that the common practice of regarding these properties as axioms, and of comparing different notions of independence by means of the graphoid axioms that they satisfy, is flawed. Of course, when they are satisfied, graphoid properties are important and useful. However, one should be very careful in regarding them as axioms. For example, if we were to impose on epistemic h-irrelevance an asymmetric version of the 'axiom' of contraction, it would require that

$$\big(\mathrm{HIR}(I, S \mid C) \text{ and } \mathrm{HIR}(I, W \mid C \cup S)\big) \Rightarrow \mathrm{HIR}(I, S \cup W \mid C) \tag{8}$$

If we choose $C = \emptyset$ here, then basically, this property requires that if X_I is irrelevant to our beliefs about X_S and irrelevant to our conditional beliefs about X_W given X_S, then X_I should also be irrelevant to our joint beliefs about $X_{S \cup W}$. I do not consider it reasonable to enforce this, because essentially, it requires that our beliefs about $X_{S \cup W}$ should be completely determined by our beliefs about X_S and our conditional beliefs about X_W given X_S. For probabilities, this is trivially true—under strict positivity assumptions—because it follows from Bayes's rule. However, for more general belief models, such as sets of probabilities, it is well known that this is not always the case. I think that this is perfectly normal, and that there is no fundamental reason why such a property should be enforced. For that reason, I consider it unreasonable to regard contraction as an axiom, at least not in general. A similar argument can be used to question the axiomatic status of the intersection property.

7. Conclusions and a Brief Look Beyond the Horizon

This paper has established two things. On the one hand, it has shown that full conditional probability measures can be used to develop a rigorous yet simple definition of a credal network under epistemic irrelevance and its irrelevant natural extension. On the other hand, it has shown that such a network satisfies many powerful theoretical properties, which can be used to develop efficient computational methods that decompose large optimisation problems into multiple smaller ones.

Using these results as a starting point, the first important future step is now to develop efficient algorithms for credal networks whose graph is more general than a tree. For some types of graphs—including but not limited to trees—examples of such algorithms can already be found in (De Bock, 2015, Chapter 7). However, for other types of graphs, the search for efficient algorithms remains open. Since the existing algorithms all focus on exact computations, approximate algorithms would definitely be worth exploring too.

The next step would then be to apply credal networks under epistemic irrelevance to real applications, in situations where the defining assumptions of a Bayesian network—exactly specified probabilities and exact independence assessments—are unrealistic. In principle, this is already feasible now: the algorithms in (De Bock, 2015) should already allow practitioners to solve large classes of problems that are relevant to their applications. However, in practice, two additional steps are needed. First of all, it is necessary to implement existing and/or new algorithms, and to develop user-friendly software to compute with them; no such software currently exists. Secondly, it should be thoroughly tested whether the bounds that are computed by these algorithms are informative enough to be useful in practice. Since epistemic irrelevance imposes less stringent constraints than complete or strong independence, the bounds of a credal network under epistemic irrelevance will be more conservative than those that correspond to other types of credal networks, and possibly too conservative to be of practical use. Although this type of behaviour does not occur in the proofs of concept in (De Bock and de Cooman, 2014; Benavoli et al., 2011), it remains to be seen whether this will be the case in other applications as well.

Acknowledgments

I am a Postdoctoral Fellow of the Research Foundation - Flanders (FWO) and wish to acknowledge its financial support. I also thank the reviewers for their detailed reading and constructive feedback.

References

A. Antonucci, C. P. de Campos, and M. Zaffalon. Probabilistic graphical models. In *Introduction to Imprecise Probabilities*, pages 207–229. Wiley, 2014.

A. Benavoli, M. Zaffalon, and E. Miranda. Robust filtering through coherent lower previsions. *IEEE Transactions on Automatic Control*, 56(7):1567–1581, 2011.

F. G. Cozman. Irrelevance and independence relations in quasi-bayesian networks. In *Proceedings of UAI '98*, pages 89–96, San Francisco, 1998.

F. G. Cozman. Credal networks. *Artificial Intelligence*, 120(2):199–233, 2000.

F. G. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39(2):167–184, 2005.

F. G. Cozman. Independence for sets of full conditional probabilities, sets of lexicographic probabilities, and sets of desirable gambles. In *Proceedings of ISIPTA '13*, pages 87–97. Compiègne, 2013.

F. G. Cozman and P. Walley. Graphoid properties of epistemic irrelevance and independence. *Annals of Mathematics and Artificial Intelligence*, 45(1-2):173–195, 2005.

J. De Bock. *Credal networks under epistemic irrelevance: theory and algorithms*. PhD thesis, 2015. URL `http://hdl.handle.net/1854/LU-6907551`.

J. De Bock and G. de Cooman. Credal networks under epistemic irrelevance using sets of desirable gambles. In *Proceedings of ISIPTA '13*, pages 99–108. SIPTA, Compiègne, 2013.

J. De Bock and G. de Cooman. An efficient algorithm for estimating state sequences in imprecise hidden Markov models. *Journal of Artificial Intelligence Research*, 50:189–233, 2014.

J. De Bock and G. de Cooman. Credal networks under epistemic irrelevance: the sets of desirable gambles approach. *International Journal of Approximate Reasoning*, 56:178–207, 2015.

C. P. de Campos and F. G. Cozman. Computing lower and upper expectations under epistemic independence. *International Journal of Approximate Reasoning*, 44(3):244–260, 2007.

G. de Cooman, F. Hermans, A. Antonucci, and M. Zaffalon. Epistemic irrelevance in credal nets: the case of imprecise Markov trees. *International Journal of Approximate Reasoning*, 51(9): 1029–1052, 2010.

L. E. Dubins. Finitely additive conditional probabilities, conglomerability and disintegrations. *The Annals of Probability*, 3(1):89–99, 1975.

E. Fagiuoli and M. Zaffalon. 2U: an exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106(1):77–107, 1998.

D. D. Mauá, C. P. de Campos, A. Benavoli, and A. Antonucci. Probabilistic inference in credal networks: new complexity results. *Journal of Artificial Intelligence Research*, 50:603–637, 2014.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Mateo, 1988.