Bayesian Networks as Coalgebras

Bart Jacobs

Radboud University, Nijmegen, The Netherlands
bart@cs.ru.nl

Abstract

A coalgebraic description of (discrete) Bayesian networks is presented. The coalgebra map representing a network sends a node to its set of predecessors, together with the associated conditional probability tables. We use this description to describe the semantics of a network in terms of various (discrete) probability distributions associated with a node: local, joint, and conditional distributions. In the background simple Python scripts are used to compute these distributions. The local and joint distributions are defined 'recursively', following the coalgebra structure. Underlying this approach are some basic properties of the (discrete probability) distribution monad. In the end we identify some new structure of the distribution monad and isolate it in what we call a 'relatively monoidal' functor.

1998 ACM Subject Classification F.1.1 Models of Computation

Keywords and phrases Bayesian network, coalgebra, probability distribution

1 Introduction

This paper attempts to connect two active areas of research, namely the (huge) area of Bayesian networks and the (much smaller) area of coalgebra. A Bayesian network is a graphical model, see [9], that describes probabilistic conditional dependencies; such networks are used for inference and (machine) learning in many, many applications these days. A coalgebra is a mathematical abstraction for a state-based dynamical system, given by a state space together with a transition function that sends states to their successors. The papers shows:

- how to capture a (discrete) Bayesian network in coalgebraic terms, where the nodes form the states, and the transition function sends a node to its set of predecessor nodes together with an associated conditional probability table;
- how to systematically develop the semantical basis of Bayesian networks in terms of (discrete) probability distributions, starting from this coalgebraic represenation.

Coalgebra has become a popular formalism in theoretical computer science for abstractly describing behaviour of state-based systems, bisimulation (observational indistinguishability) of states, and modal logics. The mathematical basis for coalgebras is provided by category theory [7, 1, **?**]. This is a branch of mathematics that is indispensable nowadays in the semantics of programming languages. Category theory provides a language of 'objects and arrows' that emphasises the structural similarities between various mathematical structures.

The area of coalgebra heavily uses the categorical notions of functor and monad. These functors/monads F are used to capture different sorts of computation, via maps (coalgebras) of the form $X \to F(X)$, where X is the state space, and the arrow describes the transition function. For instance, non-deterministic computation is captured by maps of the form $c\colon X \to \mathcal{P}(X)$, where the transition function c sends a state $x \in X$ to a *subset* of successor states $c(x) \subseteq X$. Partial computation is described via coalgebra $c\colon X \to \{\bot\} \cup X$, where $c(x) = \bot$ represents a failed computation with no successor state. In the current context the


Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

distribution monad $\mathcal{D}$ is most relevant. It is used to model discrete probabilistic computation: coalgebras $X \to \mathcal{D}(X)$ may be identified with (discrete) Markov chains.

This coalgebraic approach is to a large extend 'modular' in the functor/monad F involved. For instance, there is also a 'Giry' monad $\mathcal{G}$ that captures continuous probabilistic computation. By using $\mathcal{G}$ instead of $\mathcal{D}$, one smoothly moves from discrete to continuous probabilistic systems. The levels of abstraction and modularity provided by the theory of coalgebras are useful to see what the essentials are of the constructions at hand, and how to generalise them to a different setting (like quantum Bayesian networks in [5]). This categorical analysis of Bayesian networks is still in its infancy. One other source is [3], where continuous systems and the Giry monad $\mathcal{G}$ (but no coalgebras) are used instead of discrete systems and the distribution monad $\mathcal{D}$ that are used here.

In the theory of Bayesian networks one distinguishes learning and inference. Roughly, learning involves finding out the graphical structure, by analysing what the (in)dependencies are within a big joint probability distribution covering all the nodes. This is not what we do here. We assume that the graphical structure is already given (somehow), and we use its coalgebraic representation to deduce additional information in terms of probability distributions associated to individual nodes. Via such distributions one can do inference: updating distributions in the light of certain evidence.

The paper is organised as follows. It starts with notation, basic operations, and definitions for distributions and predicates. Subsequently, Section 3 introduces the coalgebraic represenation of Bayesian networks. This section is rather concrete and elaborates an example. We do this in great detail because we think that the best way to make people from different communities see what is going on is to focus on concrete examples. However, this section contains one simple, general construction, namely the 'local' distribution associated with each node in a Bayesian network (as coalgebra). Next, Section 4 describes how to obtain a 'joint' distribution for each node, depending on its predecessor nodes. This is rather subtle for the case when a node has multiple predecessors which again have common predecessors (like in a diamond $\diamondsuit$ in the graph). Once these joint distributions are in place, conditional distributions are described in Section 5; they can be used for actual inference. Finally, Section 6 is meant for the categorically interested reader: it abstracts some structure of the distribution monad $\mathcal{D}$ that we exploited in order to define joint distributions (in the diamond case) into a new notion that we call 'relatively monoidal'.

2 Preliminaries on predicates and distributions

A *fuzzy predicate*, or simply a predicate, on a set X is a function $p\colon X \to [0,1]$, where $[0,1] \subseteq \mathbb{R}$ is the unit interval. One can read $p(x) \in [0,1]$ as the degree of truth, or as the probability, of $x \in X$. We write $[0,1]^X$ for the set of predicates on X. For $p \in [0,1]^X$ there is an orthocomplement $p^\perp \in [0,1]^X$ of p, given by $p^\perp(x) = 1 - p(x)$. It is probabilistic negation, and satisfies $p^{\perp\perp} = p$.

A (finite, discrete probability) *distribution* on a set X is a formal convex sum:

$$r_1|x_1\rangle + \cdots + r_n|x_n\rangle \qquad \text{where} \qquad \begin{cases} r_1, \ldots, r_n \in [0,1] \text{ with } \sum_i r_i = 1 \\ x_1, \ldots, x_n \in X. \end{cases}$$

The 'ket' notation $|-\rangle$ is just syntactic sugar, used to distinguish $x \in X$ from its occurrence in such a sum. One can read the $r_i \in [0,1]$ as the probability that $x_i \in X$ occurs. We write $\mathcal{D}(X)$ for the set of all such distributions on X. A distribution on X can equivalently

be described as a function $\varphi\colon X \to [0,1]$ with finite support and with $\sum_x \varphi(x) = 1$. The support is the set of elements $x \in X$ with $\varphi(x) \neq 0$. A distribution is often called a *state*, and may describe what we know with which level of certainty about the various options x_i.

If we write n for the n-element set $\{0, 1, 2, \ldots, n-1\}$, then we find $\mathcal{D}(0) = 0$, $\mathcal{D}(1) = 1$, and $\mathcal{D}(2) \cong [0,1]$. The latter holds because a distribution $r|0\rangle + (1-r)|1\rangle$ in $\mathcal{D}(2)$ is completely determined by the number $r \in [0,1]$.

The assignment $X \mapsto \mathcal{D}(X)$ forms a *monad*, but we don't really need this notion for the time being. We can work with the relevant operations directly. The singleton (Dirac) distributions are given by a function $\eta\colon X \to \mathcal{D}(X)$ via $\eta(x) = 1|x\rangle$. For a function $f\colon X \to \mathcal{D}(Y)$ there is an an associated *Kleisli extension* map $f_*\colon \mathcal{D}(X) \to \mathcal{D}(Y)$, given by:

$$f_*\big(r_1|x_1\rangle + \cdots + r_n|x_n\rangle\big) = \textstyle\sum_i r_i f(x_i). \tag{1}$$

The right-hand-side is a convex sum of distributions, obtained by multiplying all the probabilities in the distribution $f(x_i)$ with r_i. This yields a distribution again.

Another fundamental operation on distributions is the merge, given by:

$$\begin{array}{rcl} \mathcal{D}(X_1) \times \cdots \times \mathcal{D}(X_n) & \xrightarrow{\quad merge \quad} & \mathcal{D}(X_1 \times \cdots \times X_n) \\ (\varphi_1, \ldots, \varphi_n) & \longmapsto & \lambda(x_1, \ldots, x_n).\, \varphi_1(x_1) \cdot \ldots \cdot \varphi_n(x_n) \end{array} \tag{2}$$

There are marginalisation maps that decompose distributions on a product into separate distributions, in the following way. For each $1 \leq i \leq n$ we have:

$$\mathcal{D}(X_1 \times \cdots \times X_n) \xrightarrow{marge_i} \mathcal{D}(X_i) \qquad \varphi \longmapsto \lambda x_i \in X_i. \sum_{j \neq i, x_j \in X_j} \varphi(x_1, \ldots, x_n) \tag{3}$$

It is not hard to see that $marge_i\big(merge(\varphi_1, \ldots, \varphi_n)\big) = \varphi_i$. These *merge* and *marge* maps are useful for expressing independence: a distribution $\varphi \in \mathcal{D}(X_1 \times \cdots \times X_1)$ is independent in the X_i if φ is the 'merge of its marginals', as expressed by the equation $\varphi = merge\big(marge_1(\varphi), \ldots, marge_n(\varphi)\big)$.

Categorically, marginalisation is described as $marge_i = \mathcal{D}(\pi_i) = (\eta \circ \pi_i)_*$, using that $\mathcal{D}$ is a functor that can be applied to the projection $\pi_i\colon X_1 \times \cdots X_n \to X_i$. The *merge* map exists because $\mathcal{D}$ is a monoidal (aka. commutative) monad, see also Section 6 below.

Having seen predicates and distributions separately, we continue with how they interact. For a predicate $p \in [0,1]^X$ and a distribution/state $\varphi \in \mathcal{D}(X)$ we write $\varphi \models p \in [0,1]$ for the measure of truth of p in φ. This validity probability is defined as:

$$\varphi \models p = \textstyle\sum_x \varphi(x) \cdot p(x) \in [0,1]. \tag{4}$$

If $\varphi \models p$ is non-zero, we can form the conditional distribution $\varphi|p \in \mathcal{D}(X)$ (like in [4]). It is pronounced as 'φ, given p', and defined by:

$$\varphi|p = \sum\nolimits_x \tfrac{\varphi(x) \cdot p(x)}{\varphi \models p} |x\rangle. \tag{5}$$

For these conditional distributions there is the following analogue of Bayes' rule.

$$\varphi|p \models q = \frac{(\varphi|q \models p) \cdot (\varphi \models q)}{\varphi \models p} = \frac{\varphi \models p \wedge q}{\varphi \models p} \qquad \text{where} \quad (p \wedge q)(x) = p(x) \cdot q(x).$$

In traditional approaches the rule of Bayes is used to calculate individual probabilities. In contrast, here we calculate with distributions, involving all these probabilities together. Our distribution-based approach is computationally less efficient — because we may calculate too much — but it is semantically clearer, as we claim.

▶ Notation 1. In the sequel we use the following notational convention. Let U be a finite set, say for instance $U = \{a, b, c, d\}$. A subset $V \subseteq U$ is then often identified with a sequence of members and non-members. An illustration works best to explain what we mean. We write the subset $\{a, c\} \subseteq U$ often as $ab^{\perp}cd^{\perp}$, and $\{d\} \subseteq U$ as $a^{\perp}b^{\perp}c^{\perp}d$. Each element $x \in U$ occurs in this notation either as x, if it is in the subset, or as $x^{\perp}$ if it is not. The order in these sequences is irrelevant. Thus the full subset $U \subseteq U$ is $abcd$ and the emptyset is $a^{\perp}b^{\perp}c^{\perp}d^{\perp}$.

3 Bayesian networks, as coalgebras, with their local distributions

For an arbitrary set X we write $\mathcal{P}(X)$ for the powerset of all subsets of X, and $\mathcal{P}_{\mathrm{f}}(X) \subseteq \mathcal{P}(X)$ for the set of all finite subsets of X. We define:

$$\mathcal{B}(X) = \coprod_{U \in \mathcal{P}_{\mathrm{f}}(X)} [0,1]^{\mathcal{P}(U)}. \tag{6}$$

This dependent sum $\mathcal{B}(X)$ contains pairs $\langle U, p\rangle$ where $U \subseteq X$ is a finite subset and $p\colon \mathcal{P}(U) \to [0,1]$ is a predicate on the subsets of U.

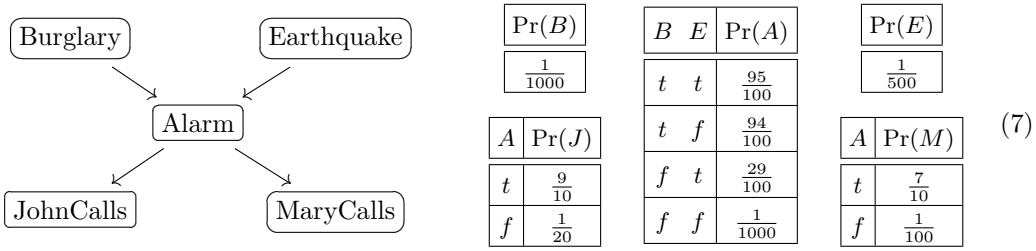
It is easy to see that $\mathcal{B}$ is a functor $\mathcal{B}\colon \mathbf{Sets} \to \mathbf{Sets}$. For a function $f\colon X \to Y$ we define $\mathcal{B}(f)\colon \mathcal{B}(X) \to \mathcal{B}(Y)$ as:

$$\mathcal{B}(f)(U, p) = (\, f(U),\ p \circ f^{-1}\,).$$

For $U \subseteq X$ we write $f(U)$ for the direct image $\mathcal{P}(f)(U) = \{f(x) \mid x \in U\} \subseteq Y$. The predicate $p\colon \mathcal{P}(U) \to [0,1]$ on $\mathcal{P}(U)$ is turned into a predicate $p \circ f^{-1}\colon \mathcal{P}(f(U)) \to [0,1]$ on $\mathcal{P}(f(U))$, given by $(V \subseteq f(U)) \mapsto p(\{x \in U \mid f(x) \in V\})$. Clearly, $\mathcal{B}$ preserves identity maps and composition.

We are going to investigate *coalgebras* of this functor $\mathcal{B}$, that is, maps of the form $X \to \mathcal{B}(X)$. We claim that Bayesian networks are such coalgebras.

We consider the classical example Bayesian network of [10, Chap. 14]. It consists of a graph together with conditional probability tables:



(7)

The alarm should be triggered by burglaries, but in practice it may also be triggered by earthquakes (with different probability). When the alarm goes off, the neighbours John and Mary may give you a call. They may also call anyway. For instance, Mary does so with probability of 1%.

In the graph on the left, each node with n predecessors, comes equipped with a predicate in $[0,1]^{2^n} = [0,1]^{\mathcal{P}(n)}$, as described via the tables on the right, where A = Alarm, B = Burglary *etc.* These tables are often read as conditional probabilities. For instance, the four entries in the table for A can also be written as $\Pr(A \mid B, E)$, $\Pr(A \mid B, E^{\perp})$, $\Pr(A \mid B^{\perp}, E)$, and $\Pr(A \mid B^{\perp}, E^{\perp})$.

Our first aim is describe this Bayesian network as an 'alarm' $\mathcal{B}$-coalgebra $a\colon N \to \mathcal{B}(N)$. The nodes of the network (7) form the set of elements of the state space N of the coalgebra. Thus we define:

$$N = \{B, E, A, J, M\}.$$

The outcome $a(x) \in \mathcal{B}(N) = \coprod_{U \in \mathcal{P}_{\mathrm{f}}(N)} [0,1]^{\mathcal{P}(U)}$ will be defined below, for each $x \in N$. An important point is that the coalgebra follows the network 'from below', describing the predecessors of a node, together with its own predicates. The (probabilistic) influence of a node's predecessors on that node does work in forward direction.

$$\begin{array}{ll} a(B) = \langle \emptyset, \langle \frac{1}{1000} \rangle \rangle & a(A) = \langle \{B, E\}, \langle \frac{95}{100}, \frac{94}{100}, \frac{29}{100}, \frac{1}{1000} \rangle \rangle \\ a(E) = \langle \emptyset, \langle \frac{1}{500} \rangle \rangle & a(J) = \langle \{A\}, \langle \frac{9}{10}, \frac{1}{20} \rangle \rangle \\ & a(M) = \langle \{A\}, \langle \frac{7}{10}, \frac{1}{100} \rangle \rangle. \end{array} \tag{8}$$

It is convenient to split the function $a = (a_1, a_2)$ into two functions, with:

$$a_1 \colon N \longrightarrow \mathcal{P}_{\mathrm{f}}(N) \qquad \text{and} \qquad a_2 \in \prod_{x \in X} [0,1]^{\mathcal{P}(a_1(x))}.$$

We see that the definition of the coalgebra in (8) closely follows the description of the Bayesian network, where the function $a_1 \colon N \to \mathcal{P}_{\mathrm{f}}(N)$ corresponds to the graph in (7) and the predicates $a_2(x) \colon \mathcal{P}(a_1(x)) \to [0,1]$, for $x \in N$, correspond to the probability tables in (7). Via this split $a = (a_1, a_2)$, and Notation 1, we can now re-write the definition of the function a on, say node $A \in N$ from (8), as:

$$a_1(A) = \{B, E\} \qquad \text{and} \qquad \begin{array}{ll} a_2(A)(BE) = \frac{95}{100} & a_2(A)(B^\perp E) = \frac{29}{100} \\ a_2(A)(BE^\perp) = \frac{94}{100} & a_2(A)(B^\perp E^\perp) = \frac{1}{1000}. \end{array}$$

We start with an elementary construction that allows us to calculate, 'in forward style', for each node x an associated 'local' distribution $\omega_x \in \mathcal{D}(2)$. It is of the form $r|x\rangle + (1-r)|x^\perp\rangle$, and describes the unconditional probability $r \in [0,1]$ that event/node x will happen.

▶ **Definition 2.** Given an arbitrary coalgebra $c \colon X \to \mathcal{B}(X)$. Let $x \in X$ be a node with n predecessors, *i.e.* with $|c_1(x)| = n$. Then it gives rise to a transformation of distributions:

$$\mathcal{D}(2)^n \xrightarrow{\overrightarrow{c(x)}} \mathcal{D}(2)$$

We call this $\overrightarrow{c(x)}$ the *forward distribution transformation*, since it combines distributions for the n predecessors of node x into a distribution for x itself. It is obtained via Kleisli extension $(-)_*$ as:

$$\overrightarrow{c(x)} = \Big(\mathcal{D}(2)^n \xrightarrow{\text{merge}} \mathcal{D}(2^n) \cong \mathcal{D}(\mathcal{P}(c_1(x))) \xrightarrow{c_2(x)_*} [0,1] \cong \mathcal{D}(2) \Big). \tag{9}$$

Let $\omega_1, \ldots, \omega_n \in \mathcal{D}(2)$ be local distributions for the predecessors, then we define:

$$\omega_x = \overrightarrow{c(x)}(\omega_1, \ldots, \omega_n) \in \mathcal{D}(2).$$

Notice that if x has no predecessors — *i.e.* if $n = |c_1(x)| = 0$ — then $\overrightarrow{c(x)} \in \mathcal{D}(2)$ is simply the distribution corresponding to the probability $c_2(x) \in [0,1]$. By starting from the initial nodes we can thus compute for each node $x \in X$ a 'local' distribution $\omega_x \in \mathcal{D}(2)$, via $\overrightarrow{c(x)}$, using the n probabilities $\omega_{x_1}, \ldots \omega_{x_n} \in \mathcal{D}(2)$ of predecessor nodes x_i in the subset $c_1(x) = \{x_1, \ldots, x_n\}$. We shall write these distributions ω_x of the form $\omega_x = r|x\rangle + (1-r)|x^\perp\rangle$, describing with probability $r \in [0,1]$ that x happens, and with probability $1-r$ that x does not happen. Notice that according to Notation 1, this ω_x is a

distribution in $\mathcal{D}(\mathcal{P}(\{x\})) \cong \mathcal{D}(2)$. Explicitly, this number r in $\omega_x = r|x\rangle + (1-r)|x^\perp\rangle$ is computed as:

$$r = \sum_{U \subseteq \{x_1,\ldots,x_n\}} c_2(x)(U) \cdot \omega_{x_i}(U \cap \{x_i\}) \tag{10}$$

Our Python script quickly computes these numbers for the different nodes as:

```
B: 0.001          A: 0.002516442          J: 0.0521389757
E: 0.002                                  M: 0.01173634498
```
(11)

We shall elaborate how these numbers arise via the forward distribution transformations $\overrightarrow{a(-)}$, for the running example coalgebra $a\colon N \to \mathcal{B}(N)$ from (8), based on the Bayesian network in (7), and how these transformations lead to local distributions $\omega_B, \omega_E, \omega_A, \omega_J, \omega_M \in \mathcal{D}(2)$ for each node. We start at the top of the graph in (7), with the nodes B and E without predecessors. The associated local distributions are simply:

$$\omega_B = \overrightarrow{a(B)} = \tfrac{1}{1000}|B\rangle + \tfrac{999}{1000}|B^\perp\rangle \quad \omega_E = \overrightarrow{a(E)} = \tfrac{1}{500}|E\rangle + \tfrac{499}{500}|E^\perp\rangle.$$

These two distributions yield, via the map $\mathit{merge}\colon \mathcal{D}(2) \times \mathcal{D}(2) \to \mathcal{D}(2 \times 2)$ from (2), a new distribution on $2 \times 2 = 4 = \{BE, BE^\perp, B^\perp E, B^\perp E^\perp\}$, namely

$$\begin{aligned}
&\mathit{merge}(\omega_B, \omega_E) \\
&= \tfrac{1}{1000} \cdot \tfrac{1}{500}|BE\rangle + \tfrac{1}{1000} \cdot \tfrac{499}{500}|BE^\perp\rangle + \tfrac{999}{1000} \cdot \tfrac{1}{500}|B^\perp E\rangle + \tfrac{999}{1000} \cdot \tfrac{499}{500}|B^\perp E^\perp\rangle \\
&= \tfrac{1}{500.000}|BE\rangle + \tfrac{499}{500,000}|BE^\perp\rangle + \tfrac{999}{500,000}|B^\perp E\rangle + \tfrac{498,501}{500,000}|B^\perp E^\perp\rangle.
\end{aligned}$$

This joint distribution clearly captures the probabilities of the yes/no possibilities for both B = Burglary and E = Earthquake.

We recall from (8) that the predicate associated with the Alarm node is $a_2(A)\colon 4 \to [0,1] \cong \mathcal{D}(2)$, given by the 4-tuple $\langle \frac{95}{100}, \frac{94}{100}, \frac{29}{100}, \frac{1}{1000}\rangle$. We apply its Kleisli extension to the above distribution on 4, as in (9), and obtain the local distribution for node A.

$$\begin{aligned}
\omega_A &= \overrightarrow{a(A)}(\omega_B, \omega_E) = a_2(A)_*\big(\mathit{merge}(\omega_B, \omega_E)\big) \\
&= \tfrac{950+469,060+289,710+498,501}{500,000,000}|A\rangle + \tfrac{50+29,940+709,290+498,002,499}{500,000,000}|A^\perp\rangle \\
&= \tfrac{1,258,221}{500,000,000}|A\rangle + \tfrac{498,741,779}{500,000,000}|A^\perp\rangle.
\end{aligned}$$

This first number $\frac{1,258,221}{500,000,000}$ equals 0.002516442 as listed in the Python output (11). It is the (non-conditional) probability that an alarm will be raised. It can also be computed as the validity probability $\mathit{merge}(\omega_B, \omega_E) \models a_2(A)$, as defined in (4).

The remaining two local distributions ω_J and ω_M for nodes J and M are easier, since for $n = 1$ predecessor the map $\mathit{merge}\colon \mathcal{D}(2) \to \mathcal{D}(2)$ is the identity and can be omitted. We get:

$$\tfrac{521,389,757}{10,000,000,000}|J\rangle + \tfrac{9,478,610,243}{10,000,000,000}|J^\perp\rangle \qquad \tfrac{586,817,249}{50,000,000,000}|M\rangle + \tfrac{49,413,182,751}{50,000,000,000}|M^\perp\rangle.$$

This last number, close to 1, is the probability $\Pr(M^\perp)$ that Mary will not call.

4 Joint distributions

In the previous section we have computed for each label/state x a 'local' distribution in $\mathcal{D}(2)$, involving only $|x\rangle$ and $|x^\perp\rangle$. In this section we associate with x another, more complicated 'joint' distribution, involving all predecessors of x, cumulatively.

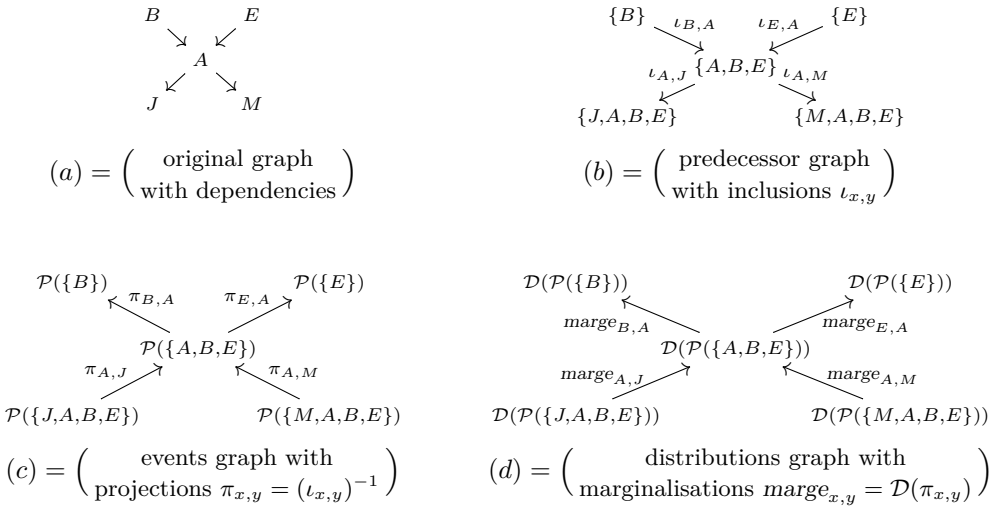


Figure 1 The original Alarm graph (a), as in (7), with three successive transformations $(a) \to (b) \to (c) \to (d)$.

Thus, for a node $x \in X$ of a coalgebra $c = (c_1, c_2) \colon X \to \mathcal{B}(X)$ we first define the subset $\mathit{precs}(x) \subseteq X$ of (cumulative) predecessor nodes via:

$$\mathit{precs}(x) \;=\; \{x\} \cup \bigcup_{x' \in c_1(x)} \mathit{precs}(x'). \tag{12}$$

This unfolding in this definition terminates if the set of nodes X is finite, and the coalgebra does not contain cycles. This is always the case for Bayesian networks. Notice that we include the node x in the subset of predecessors. In this way, for a node without predecessors, that is for $x \in X$ with $c_1(x) = \emptyset$, we get $\mathit{precs}(x) = \{x\}$.

Thus, in the alarm coalgebra of the previous section we get, as expected,

$$\begin{array}{lll} \mathit{precs}(B) = \{B\} & & \\ \mathit{precs}(E) = \{E\} & \mathit{precs}(A) = \{A, B, E\} & \mathit{precs}(J) = \{J, A, B, E\} \\ & & \mathit{precs}(M) = \{M, A, B, E\}. \end{array}$$

Our aim is to define for each node x a joint distribution on the set $\mathcal{P}(\mathit{precs}(x))$ of subsets (events) of predecessors of x. We do so by first transforming the original graph into other graphs, in three steps. These steps are described in Figure 1 for our running example. Our aim is to find joint distributions, written as χ_x, for node x, such that $\mathit{marge}_{x,y}(\chi_y) = \chi_x$ in the distributions graph (d).

We take a closer look at what is going on in Figure 1 in more general terms. If, for an arbitrary coalgebra $c \colon X \to \mathcal{B}(X)$ we have a predecessor set $c_1(x) = \{x_1, \ldots, x_n\}$, then the dependency arrows $x_i \to x$ in diagram (a) translate into inclusion arrows $\iota_{x_i,x} \colon \mathit{precs}(x_i) \hookrightarrow \{x\} \cup \bigcup_i \mathit{precs}(x_i) = \mathit{precs}(x)$ in diagram (b).

Next, the inverse image arrows $\pi_{x_i,x} = (\iota_{x_i,x})^{-1} \colon \mathcal{P}(\mathit{precs}(x)) \to \mathcal{P}(\mathit{precs}(x_i))$ in the opposite direction in diagram (c) are given by restriction:

$$\pi_{x_i,x}(U) \;=\; (\iota_{x_i,x})^{-1}(U) \;=\; \{z \in \mathit{precs}(x_i) \mid \iota_{x_i,x}(z) \in U\} \;=\; U \cap \mathit{precs}(x_i).$$

The marginalisation maps $\mathit{marge}_{x_i,x} = \mathcal{D}(\pi_{x_i,x}) \colon \mathcal{D}(\mathcal{P}(\mathit{precs}(x))) \to \mathcal{D}(\mathcal{P}(\mathit{precs}(x_i)))$ in

diagram (d) are then given by:

$$\begin{aligned} \mathit{marge}_{x_i,x}(\varphi) \;=\; \mathcal{D}(\pi_{x_i,x})(\varphi) &= \sum_{V \subseteq \mathit{precs}(x_i)} \left(\sum_{U \in \pi^{-1}_{x_i,x}(V)} \varphi(U) \right) \big| V \big\rangle \\ &= \sum_{V \subseteq \mathit{precs}(x_i)} \left(\sum_{U \subseteq \mathit{precs}(x), U \cap \mathit{precs}(x_i) = V} \varphi(U) \right) \big| V \big\rangle. \end{aligned}$$

As stated, our aim is to define joint distributions χ_x on $\mathcal{P}(\mathit{precs}(x))$. We wish to do this 'recursively', following the structure of the graph, moving from joint distributions for parents to joint distributions for children. For 'initial' nodes, without parents, the joint distribution will equal the local distribution: $\chi_x = \omega_x$.

For nodes x with parents x_i we distinguish whether these parents have disjoint predecessors or not, that is whether $\mathit{precs}(x_i) \cap \mathit{precs}(x_j) = \emptyset$ or not, for different parents x_i, x_j. The case where all predecessor sets are disjoint will be handled first. This is the case for instance for our Alarm example: the only node with multiple parents is A, with parents B, E satisfying $\mathit{precs}(B) \cap \mathit{precs}(E) = \{B\} \cap \{E\} = \emptyset$.

▶ **Definition 3.** Let $c\colon X \to \mathcal{B}(X)$ be an arbitrary coalgebra, where X is finite and $c_1\colon X \to \mathcal{P}_{\mathrm{f}}(X)$ is acyclic. For each node $x \in X$ we define a *joint distribution* $\chi_x \in \mathcal{D}\big(\mathcal{P}(\mathit{precs}(x))\big)$ in the following recursive way, following (12).

Let $x \in N$ have set of predecessors $c_1(x) = \{x_1, \ldots, x_n\} \subseteq X$, where $\mathit{precs}(x_i) \cap \mathit{precs}(x_j) = \emptyset$ for $i \neq j$. Assume that we already have joint distributions $\chi_{x_i} \in \mathcal{D}\big(\mathcal{P}(\mathit{precs}(x_i))\big)$. Then we define $\chi_x \in \mathcal{D}\big(\mathcal{P}(\mathit{precs}(x))\big)$, where $\mathit{precs}(x) = \{x\} \cup \mathit{precs}(x_1) \cup \cdots \cup \mathit{precs}(x_n)$, as:

$$\begin{aligned} \chi_x = \sum_{U \subseteq \bigcup_i \mathit{precs}(x_i)} & c_2(x)(U \cap c_1(x)) \cdot \textstyle\prod_i \chi_{x_i}(U \cap \mathit{precs}(x_i)) \big| xU \big\rangle \\ & + (1 - c_2(x)(U \cap c_1(x))) \cdot \textstyle\prod_i \chi_{x_i}(U \cap \mathit{precs}(x_i)) \big| x^{\perp} U \big\rangle \end{aligned} \tag{13}$$

For the more categorically oriented reader we sketch what the formula (13) for the joint distribution χ_x means categorically. Assume again $\mathit{precs}(x) = \{x\} \cup \bigcup_i \mathit{precs}(x_i)$. We use ad hoc notation $\mathit{pprecs}(x) = \mathit{precs}(x) = \{x\} = \bigcup_i \mathit{precs}(x_i)$ for the set of 'proper' predecessors. Since all these unions are disjoint ones, we may understand them categorically as coproducts $+$. We use that the powerset functor $\mathcal{P}$ sends coproducts to products, as in: $\mathcal{P}(A + B) \cong \mathcal{P}(A) \times \mathcal{P}(B)$. Hence we can form the composite map in Figure 2, that turns joint distributions $\chi_{x_i} \in \mathcal{D}(\mathcal{P}(\mathit{precs}(x_i)))$ for the predecessors x_i into a joint distribution $\chi_x \in \mathcal{D}(\mathcal{P}(\mathit{precs}(x)))$ for x.

The marked arrow $(*)$ in Figure 2 involves a 'Kleisli extension in context', in which a map $f\colon A \to \mathcal{D}(B)$ is extended to a map $\mathcal{D}(A \times C) \to \mathcal{D}(B \times A \times C)$, via:

$$\mathcal{D}(A \times C) \xrightarrow{\mathcal{D}(\langle f \circ \pi_1, \mathrm{id} \rangle)} \mathcal{D}(\mathcal{D}(B) \times A \times C) \xrightarrow{\mathcal{D}(\mathrm{st})} \mathcal{D}(\mathcal{D}(B \times A \times C)) \xrightarrow{\mu} \mathcal{D}(B \times A \times C),$$

where st and μ are the strength and multiplication of the distribution monad. This extension is applied to the conditional probability table $c_2(x)\colon \mathcal{P}(\{x_1, \ldots, x_n\}) \to [0,1] \cong \mathcal{D}(\mathcal{P}(\{x\}))$. Explicitly, this composite is: $\varphi \mapsto \lambda(b, a, c).\, f(a)(b) \cdot \varphi(a, c)$. We emphasise that this works because we are assuming that predecessor sets are disjoint.

We sketch the outcome for our alarm example, with coalgebra $a = (a_1, a_2)\colon N \to \mathcal{B}(N)$. For the initial nodes B and E we simply have $\chi_B = \omega_B$ and $\chi_E = \omega_E$, where the ω's describe

$$\begin{array}{c}
\mathcal{D}\Big(\mathcal{P}(precs(x_1))\Big) \times \cdots \times \mathcal{D}\Big(\mathcal{P}(precs(x_n))\Big) \\
\downarrow merge \\
\mathcal{D}\Big(\mathcal{P}(precs(x_1)) \times \cdots \times \mathcal{P}(precs(x_n))\Big) \\
\| \\
\mathcal{D}\Big(\mathcal{P}\big(\{x_1\} + pprecs(x_1)\big) \times \cdots \times \mathcal{P}\big(\{x_n\} + pprecs(x_n)\big)\Big) \\
\|\wr \\
\mathcal{D}\Big(\mathcal{P}(\{x_1\}) \times \mathcal{P}(pprecs(x_1)) \times \cdots \times \mathcal{P}(\{x_n\}) \times \mathcal{P}(pprecs(x_n))\Big) \\
\|\wr \\
\mathcal{D}\Big(\mathcal{P}(\{x_1, \ldots, x_n\}) \times \mathcal{P}(pprecs(x_1)) \times \cdots \times \mathcal{P}(pprecs(x_n))\Big) \\
\downarrow (*) \\
\mathcal{D}\Big(\mathcal{P}(\{x\}) \times \mathcal{P}(\{x_1, \ldots, x_n\}) \times \mathcal{P}(pprecs(x_1)) \times \cdots \times \mathcal{P}(pprecs(x_n))\Big) \\
\|\wr \\
\mathcal{D}\Big(\mathcal{P}(\{x\} + precs(x_1) + \cdots + precs(x_n))\Big) \\
\| \\
\mathcal{D}\big(\mathcal{P}(precs(x))\big)
\end{array}$$

Figure 2 The formula (13) explained as function composition

the local distributions from the previous section. For the alarm node A we get:

$$\begin{aligned}
&\sum_{U \subseteq \{B,E\}} a_2(A)(U \cap \{B,E\}) \cdot \chi_B(U \cap \{B\}) \cdot \chi_E(U \cap \{E\}) |AU\rangle \\
&\qquad + (1 - a_2(A)(U \cap \{B,E\})) \cdot \chi_B(U \cap \{B\}) \cdot \chi_E(U \cap \{E\}) |A^\perp U\rangle \\
&= \tfrac{950}{500{,}000{,}000}|ABE\rangle + \tfrac{50}{500{,}000{,}000}|A^\perp BE\rangle + \tfrac{469{,}060}{500{,}000{,}000}|ABE^\perp\rangle + \tfrac{29{,}940}{500{,}00{,}000}|A^\perp BE^\perp\rangle + \\
&\quad \tfrac{289{,}710}{500{,}000{,}000}|AB^\perp E\rangle + \tfrac{709{,}290}{500{,}000{,}000}|A^\perp B^\perp E\rangle + \tfrac{498{,}501}{500{,}000{,}000}|AB^\perp E^\perp\rangle + \tfrac{498{,}002{,}499}{500{,}000{,}000}|A^\perp B^\perp E^\perp\rangle.
\end{aligned}$$

Our Python script computes and prints it as follows, using $(-)'$ for $(-)^\perp$.

```
1.9e-06|ABE> + 0.00093812|ABE'> + 0.00057942|AB'E> + 0.000997002|AB'E'> +
  1e-07|A'EB> + 5.988e-05|A'BE'> + 0.00141858|B'A'E> + 0.996004998|B'A'E'>
```

The distributions χ_J and χ_M are obtained similarly; they involve 16 terms.

▶ Proposition 4. The definition χ_x in (13) is indeed a probability distribution. Moreover, it marginalises to:

1. the joint distributions χ_{x_i} of its parents x_i, via $marge_{x_i,x}(\chi_x) = \chi_{x_i}$;
2. the local distribution $\omega_x \in \mathcal{D}(\mathcal{P}(\{x\}))$, via the projection $\mathcal{P}(precs(x)) \to \mathcal{P}(\{x\})$ induced by the inclusion $\{x\} \hookrightarrow precs(x)$.

Proof. We first have to prove that in the formula (13) for the joint distribution χ_x the probabilities before the $|xU\rangle$ and $|x^\perp U\rangle$ add up to 1, for $U \subseteq \bigcup_i precs(x_i)$. For each such U, these probabilities before $|xU\rangle$ and $|x^\perp U\rangle$ add up to the product $\prod_i \chi_{x_i}(U \cap precs(x_i))$. Hence we have to prove that the sum of these terms is 1. But since the predecessor sets $precs(x_i)$ are disjoint, we can split the subset $U \subseteq \bigcup_i precs(x_i)$ into separate subsets

$U_i \subseteq \mathit{precs}(x_i)$, as in:

$$\begin{aligned}\sum_{U \subseteq \bigcup_i \mathit{precs}(x_i)} \prod_i \chi_{x_i}(U \cap \mathit{precs}(x_i)) &= \sum_{U_1 \subseteq \mathit{precs}(x_1)} \cdots \sum_{U_n \subseteq \mathit{precs}(x_n)} \prod_i \chi_{x_i}(U_i) \\ &= \prod_i \sum_{U_i \subseteq \mathit{precs}(x_i)} \chi_{x_i}(U_i) \\ &= \prod_i 1 \qquad \text{since each } \chi_i \text{ is a distribution} \\ &= 1.\end{aligned}$$

We only prove point (1). Via the same splitting we obtain the equation $\mathit{marge}_{x_j,x}(\chi_x) = \chi_{x_j}$ for the j-th parent. Indeed, for $V \subseteq \mathit{precs}(x_j)$ we have:

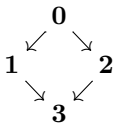
$$\begin{aligned}\mathit{marge}_{x_j,x}(\chi_x)(V) &= \sum_{U \subseteq \mathit{precs}(x), U \cap \mathit{precs}(x_j) = V} \chi_x(U) \\ &\overset{(13)}{=} \sum_{U \subseteq \bigcup_i \mathit{precs}(x_i), U \cap \mathit{precs}(x_j) = V} \prod_i \chi_{x_i}(U \cap \mathit{precs}(x_i)) \\ &= \sum_{j \neq i} \sum_{U_i \subseteq \mathit{precs}(x_i)} \chi_{x_j}(V) \cdot \prod_{i \neq j} \chi_{x_i}(U_i) \\ &= \chi_{x_j}(V) \cdot \prod_{i \neq j} \sum_{U_i \subseteq \mathit{precs}(x_i)} \chi_{x_i}(U_i) \\ &= \chi_{x_j}(V) \cdot \prod_{i \neq j} 1 \\ &= \chi_{x_j}(V).\end{aligned}$$

◀

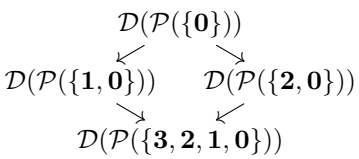
4.1 The case of non-disjoint predecessor sets

When predecessor sets are non-disjoint the situation is more complicated. We shall not attempt to give a completely general treatment, since it depends on the topology of the network. Instead, we illustrate the situation with a simple example. Subsequently, we give a more systematic description.

▶ **Example 5.** We consider another illustration for the sole purpose of demonstrating what happens when a node has predecessors with non-disjoint sets of predecessors, like in the graph on the left, with set of nodes $4 = \{\mathbf{0}, \mathbf{1}, \mathbf{2}, \mathbf{3}\}$.



with



On the right we have written the distributions graph, using the associated predecessor sets: $\mathit{precs}(\mathbf{0}) = \{\mathbf{0}\}$, $\mathit{precs}(\mathbf{1}) = \{\mathbf{1}, \mathbf{0}\}$, $\mathit{precs}(\mathbf{2}) = \{\mathbf{2}, \mathbf{0}\}$, $\mathit{precs}(\mathbf{3}) = \{\mathbf{3}, \mathbf{2}, \mathbf{1}, \mathbf{0}\}$.

We turn this diamond graph into a Bayesian network via a coalgebra $d\colon 4 \to \mathcal{B}(4)$.

$$d(\mathbf{1}) = \langle\, \{\mathbf{0}\},\ \begin{cases} \mathbf{0} \mapsto \frac{1}{3} \\ \mathbf{0}^\perp \mapsto \frac{4}{5} \end{cases} \rangle \qquad d(\mathbf{0}) = \langle\, \emptyset,\ \tfrac{1}{4} \,\rangle$$

$$d(\mathbf{2}) = \langle\, \{\mathbf{0}\},\ \begin{cases} \mathbf{0} \mapsto \frac{1}{5} \\ \mathbf{0}^\perp \mapsto \frac{1}{6} \end{cases} \rangle \qquad d(\mathbf{3}) = \langle\, \{\mathbf{1}, \mathbf{2}\},\ \begin{cases} \mathbf{1}, \mathbf{2} \mapsto \frac{1}{8} \\ \mathbf{1}, \mathbf{2}^\perp \mapsto \frac{3}{8} \\ \mathbf{1}^\perp, \mathbf{2} \mapsto \frac{5}{8} \\ \mathbf{1}^\perp, \mathbf{2}^\perp \mapsto \frac{7}{8} \end{cases} \rangle$$

The joint probabilities for the nodes $\mathbf{0}$, $\mathbf{1}$ and $\mathbf{2}$ are easy: $\chi_{\mathbf{0}} = \frac{1}{4}|\mathbf{0}\rangle + \frac{3}{4}|\mathbf{0}^{\perp}\rangle$ and:

$$\begin{aligned}\chi_{\mathbf{1}} &= \tfrac{1}{12}|\mathbf{10}\rangle + \tfrac{2}{12}|\mathbf{1}^{\perp}\mathbf{0}\rangle + \tfrac{12}{20}|\mathbf{10}^{\perp}\rangle + \tfrac{3}{20}|\mathbf{1}^{\perp}\mathbf{0}^{\perp}\rangle \\ \chi_{\mathbf{2}} &= \tfrac{1}{20}|\mathbf{20}\rangle + \tfrac{4}{20}|\mathbf{2}^{\perp}\mathbf{0}\rangle + \tfrac{3}{24}|\mathbf{20}^{\perp}\rangle + \tfrac{15}{24}|\mathbf{2}^{\perp}\mathbf{0}^{\perp}\rangle.\end{aligned}$$

The question is how to compute the joint distribution $\chi_{\mathbf{3}}$. Earlier, we would just merge the distributions $\chi_{\mathbf{1}}$ and $\chi_{\mathbf{2}}$, but in the present situation with a common ancestor we first have to form the 'conditional merge' $\chi_{\mathbf{12|0}} \in \mathcal{D}(\mathcal{P}(\{\mathbf{2},\mathbf{1},\mathbf{0}\}))$. Then we can form $\chi_{\mathbf{3}} \in \mathcal{D}(\mathcal{P}(\{\mathbf{3},\mathbf{2},\mathbf{1},\mathbf{0}\}))$ via the conditional probability table of node $\mathbf{3}$.

This distribution $\chi_{\mathbf{12|0}} \in \mathcal{D}(\mathcal{P}(\{\mathbf{0},\mathbf{1},\mathbf{2}\}))$ can be expressed 'recursively' in terms of its predecessor distributions:

$$\begin{aligned}\chi_{\mathbf{12|0}} &= \sum_{U \subseteq \{\mathbf{2},\mathbf{1}\}} \frac{\chi_{\mathbf{2}}((U \cap \mathit{precs}(2))\mathbf{0}) \cdot \chi_{\mathbf{1}}((U \cap \mathit{precs}(1))\mathbf{0})}{\chi_{\mathbf{0}}(\mathbf{0})} \big| U\mathbf{0}\big\rangle \\ &\qquad + \frac{\chi_{\mathbf{2}}((U \cap \mathit{precs}(2))\mathbf{0}^{\perp}) \cdot \chi_{\mathbf{1}}((U \cap \mathit{precs}(1))\mathbf{0}^{\perp})}{\chi_{\mathbf{0}}(\mathbf{0}^{\perp})} \big| U\mathbf{0}^{\perp}\big\rangle \\ &= \tfrac{2}{120}|\mathbf{210}\rangle + \tfrac{12}{120}|\mathbf{210}^{\perp}\rangle + \tfrac{4}{120}|\mathbf{21}^{\perp}\mathbf{0}\rangle + \tfrac{3}{120}|\mathbf{21}^{\perp}\mathbf{0}^{\perp}\rangle \\ &\qquad + \tfrac{8}{120}|\mathbf{2}^{\perp}\mathbf{10}\rangle + \tfrac{60}{120}|\mathbf{2}^{\perp}\mathbf{10}^{\perp}\rangle + \tfrac{16}{120}|\mathbf{2}^{\perp}\mathbf{1}^{\perp}\mathbf{0}\rangle + \tfrac{15}{120}|\mathbf{2}^{\perp}\mathbf{1}^{\perp}\mathbf{0}^{\perp}\rangle.\end{aligned}$$

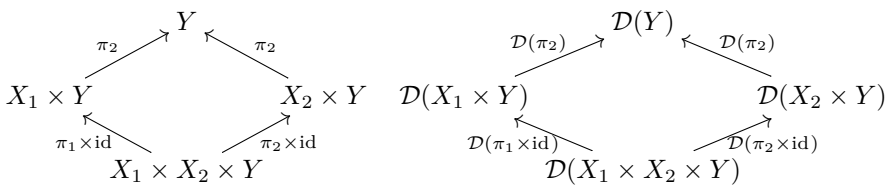
Crucially, we divide by $\chi_{\mathbf{0}}$ since it is a common predecessor, see also (14) below.

We can now calculate the joint distribution $\chi_{\mathbf{3}} \in \mathcal{D}\big(\mathcal{P}(\{\mathbf{3},\mathbf{2},\mathbf{1},\mathbf{0}\})\big)$ for the node $\mathbf{3}$ by using the predicate $d_2(\mathbf{3})\colon \mathcal{P}(\{\mathbf{1},\mathbf{2}\}) \to [0,1]$ in context. It yields:

$$\begin{aligned}\chi_{\mathbf{3}} &= \sum_{U \subseteq \{\mathbf{2},\mathbf{1},\mathbf{0}\}} d_2(\mathbf{3})(U \cap d_1(\mathbf{3})) \cdot \chi_{\mathbf{12|0}}(U)|\mathbf{3}U\rangle \\ &\qquad + (1 - d_2(\mathbf{3})(U \cap d_1(\mathbf{3}))) \cdot \chi_{\mathbf{12|0}}(U)|\mathbf{3}^{\perp}U\rangle \\ &= \tfrac{2}{960}|\mathbf{3210}\rangle + \tfrac{14}{960}|\mathbf{3}^{\perp}\mathbf{210}\rangle + \tfrac{12}{960}|\mathbf{3210}^{\perp}\rangle + \tfrac{84}{960}|\mathbf{3}^{\perp}\mathbf{210}^{\perp}\rangle \\ &\qquad + \tfrac{20}{960}|\mathbf{321}^{\perp}\mathbf{0}\rangle + \tfrac{12}{960}|\mathbf{3}^{\perp}\mathbf{21}^{\perp}\mathbf{0}\rangle + \tfrac{15}{960}|\mathbf{321}^{\perp}\mathbf{0}^{\perp}\rangle + \tfrac{9}{960}|\mathbf{3}^{\perp}\mathbf{21}^{\perp}\mathbf{0}^{\perp}\rangle \\ &\qquad + \tfrac{24}{960}|\mathbf{32}^{\perp}\mathbf{10}\rangle + \tfrac{40}{960}|\mathbf{3}^{\perp}\mathbf{2}^{\perp}\mathbf{10}\rangle + \tfrac{180}{960}|\mathbf{32}^{\perp}\mathbf{10}^{\perp}\rangle + \tfrac{300}{960}|\mathbf{3}^{\perp}\mathbf{2}^{\perp}\mathbf{10}^{\perp}\rangle \\ &\qquad + \tfrac{112}{960}|\mathbf{32}^{\perp}\mathbf{1}^{\perp}\mathbf{0}\rangle + \tfrac{16}{960}|\mathbf{3}^{\perp}\mathbf{2}^{\perp}\mathbf{1}^{\perp}\mathbf{0}\rangle + \tfrac{105}{960}|\mathbf{32}^{\perp}\mathbf{1}^{\perp}\mathbf{0}^{\perp}\rangle + \tfrac{15}{960}|\mathbf{3}^{\perp}\mathbf{2}^{\perp}\mathbf{1}^{\perp}\mathbf{0}^{\perp}\rangle.\end{aligned}$$

(Our Python script can produce it automatically.)

The crucial step is the intermediate distribution $\chi_{\mathbf{12|0}}$. We describe diagrammatically what is going on. Consider for arbitrary sets X_1, X_2, Y the projections $\pi_2\colon X_i \times Y \to Y$ on the left below, with the resulting pullback square.



The diagram on the right is obtained by applying the functor $\mathcal{D}(-)$ to the one on the left. Translated to this setting our question is: given distributions $\varphi_i \in \mathcal{D}(X_i \times Y)$ with the same Y-marginal, that is with $\mathcal{D}(\pi_2)(\varphi_1) = \varphi = \mathcal{D}(\pi_2)(\varphi_2)$, can we find an appropriate distribution $\psi \in \mathcal{D}(X_1 \times X_2 \times Y)$ which marginalises to both φ_1, φ_2?

The distribution functor $\mathcal{D}$ preserves weak pullbacks (see [8, 11]), so that the square on the right is a weak pullback. This guarantees the existence of some ψ. But that is *not* good enough. We need a canonical choice, namely:

$$\psi(x_1, x_2, y) = \frac{\varphi_1(x_1, y) \cdot \varphi_2(x_2, y)}{\varphi(y)} \tag{14}$$

Notice that if $\varphi(y) = 0$, then both $\varphi_1(x_1, y) = 0$ and $\varphi_2(x_2, y) = 0$, since $\varphi(y) = \mathcal{D}(\pi_2)(\varphi_1)(y) = \sum_x \varphi_1(x, y)$ and also $\varphi(y) = \mathcal{D}(\pi_2)(\varphi_2)(y) = \sum_x \varphi_2(x, y)$. In that case we choose 0 as value for $\psi(x_1, x_2, y)$.

It is easy to prove that this ψ in (14) marginalises to φ_1 — and similarly to φ_2.

$$\begin{aligned}\mathcal{D}(\pi_1 \times \mathrm{id})(\psi)(x_1, y) = \sum_{x_2} \psi(x_1, x_2, y) = \sum_{x_2} \frac{\varphi_1(x_1,y)\cdot\varphi_2(x_2,y)}{\varphi(y)} &= \frac{\varphi_1(x_1,y)\cdot\sum_{x_2}\varphi_2(x_2,y)}{\varphi(y)} \\ &= \frac{\varphi_1(x_1,y)\cdot\varphi(y)}{\varphi(y)} \\ &= \varphi_1(x_1, y).\end{aligned}$$

In a similar way one proves that ψ is a distribution:

$$\sum_{x_1,x_2,y} \psi(x_1, x_2, y) = \sum_y \frac{\big(\sum_{x_1} \varphi_1(x_1,y)\big)\cdot\big(\sum_{x_2} \varphi_2(x_2,y)\big)}{\varphi(y)} = \sum_y \frac{\varphi(y)\cdot\varphi(y)}{\varphi(y)} = \sum_y \varphi(y) = 1.$$

We elaborate on this construction in Section 6, using categorical language.

5 Conditional distributions

Like before we elaborate an example, but at the same time indicate what the general mechanism is. Suppose we wish, in our alarm example, to compute the conditional distribution capturing the probability of a Burglary, given that the alarm has sounded. This is done in the following two steps.

(1) **Calculate the appropriate conditional of the joint distribution**. In the previous section we have calculated the joint distribution χ_A at node A. It is a distribution on the set $\mathcal{P}(\mathit{precs}(A))$ of subsets of predecessors $\mathit{precs}(A) = \{A, B, E\}$. On this powerset $\mathcal{P}(\mathit{precs}(A))$ we consider the (fuzzy) predicate p_A describing that the alarm was raised. This predicate p_A is a function:

$$\mathcal{P}(\mathit{precs}(A)) \xrightarrow{p_A} [0, 1] \qquad \text{given by} \qquad p_A(U) = \begin{cases} 1 \text{ if } A \in U \\ 0 \text{ if } A \notin U. \end{cases}$$

We can now use (5) to calculate the conditional distribution $\chi_A|p_A$ on $\mathcal{P}(\{A, B, E\})$. First, the validity probability is:

$$\chi_A \models p_A = \sum_{U \subseteq \{A,B,E\}} \chi_A(U) \cdot p_A(U) = \sum_{U \subseteq \{B,E\}} \chi_A(AU) = \tfrac{1.258.221}{500.000.000}.$$

This is (of course) the same number that appears in the local distribution ω_A, see also Proposition 4. We can now calculate:

$$\chi_A|p_A = \sum_{U \subseteq \{A,B,E\}} \tfrac{\chi_A(U)\cdot p_A(U)}{\chi_A \models p_A} |U\rangle = \sum_{U \subseteq \{B,E\}} \tfrac{\chi_A(AU)}{\chi_A \models p_A} |AU\rangle =$$
$$\tfrac{950}{1.258.221}|ABE\rangle + \tfrac{469.060}{1.258.221}|ABE^\perp\rangle + \tfrac{289.710}{1.258.221}|AB^\perp E\rangle + \tfrac{498.501}{1.258.221}|AB^\perp E^\perp\rangle$$

These four numbers describe the conditional probabilities $\Pr(B, E|A)$, $\Pr(B, E^\perp|A)$, $\Pr(B^\perp, E|A)$, $\Pr(B^\perp, E^\perp|A)$, respectively, in a single distribution.

(2) **Calculate the appropriate marginal**. We obtain the required distribution, $\varphi =$ 'burglary, given alarm', in $\mathcal{D}(\mathcal{P}(\{B\}))$ via marginalisation, using $\pi_B \colon \mathcal{P}(\{A, B, E\}) \to \mathcal{P}(\{B\})$ given by $\pi_B(U) = U \cap \{B\}$. Thus:

$$\begin{aligned}\varphi = \mathcal{D}(\pi_B)(\chi_A|p_A) &= \sum_{U\subseteq\{B\}} \Big(\sum_{V\in\pi_B^{-1}(U)} \chi_A|p_A(V)\Big)|U\rangle \\ &= \tfrac{950+469,060}{1,258,221}|B\rangle + \tfrac{289,710+498,501}{1,258,221}|B^\perp\rangle = \tfrac{470,010}{1,258,221}|B\rangle + \tfrac{788,211}{1,258,221}|B^\perp\rangle.\end{aligned}$$

This first probability $\Pr(B|A)$, of a burglary if the alarm sounds, is roughly $0,37$. In the traditional way this same number is computed as:

$$P(B|A) = \frac{P(A,B)}{P(A)} = \frac{P(A|B,E)\cdot P(B)\cdot P(E) + P(A|B,E^\perp)\cdot P(B)\cdot P(E^\perp)}{P(A)}$$

In this example we started from the predicate p_A. It is a 'Boolean' predicate, since its outcomes are in the subset $\{0, 1\} \subseteq [0, 1]$. The approach would work equally well if we would start from a properly fuzzy predicate, say $p_A^{\frac{1}{3}}$ given by:

$$p_A^{\frac{1}{3}}(U) = \begin{cases} \frac{1}{3} \text{ if } A \in U \\ \frac{2}{3} \text{ if } A \notin U. \end{cases}$$

We would then compute the distribution 'Burglary, given that the alarm was raised with probability one third' as $\mathcal{D}(\pi_B)(\chi_A|p_A^{\frac{1}{3}})$. This is less straightforward in traditional, Bayesian approaches.

With these conditional distributions in place one can use a Bayesian network for inference. Here is a small illustration for the Alarm network. Recall from (7) that the initial/prior probabilities for burglary (B) and earthquake (E) are $\frac{1}{1000}$ and $\frac{1}{500}$. But given that John calls, we can update these probabilities, via the conditional distributions 'burglary, given John calls' and 'earthquake, given John calls' (and similarly for Mary). The following Python output describes these successively updated probabilities after a sequence of 'evidence' J, M, M, J, starting from $\frac{1}{1000}$ and $\frac{1}{500}$.

```
Given J, P(B) becomes 0.0162837299468 and P(E) becomes 0.0113949687738
Given M, P(B) becomes 0.456862955199  and P(E) becomes 0.105575234095
Given M, P(B) becomes 0.945862958301  and P(E) becomes 0.138624565179
Given J, P(B) becomes 0.994316646564  and P(E) becomes 0.141775105645
```

We see that a burglary has become the most likely event after the last call, because a burglary is more likely (than an earthquake) to lead to an alarm in (7), and thus to a call.

6 Relatively monoidal functors

This section is of a different nature. It abstracts away from Bayesian networks and tries to capture categorically the property that we used to obtain the intermediate distribution $\chi_{\mathbf{21}|\mathbf{0}}$ in Example 5. We assume familiarity with the notion of 'monoidal functor', see *e.g.* [2]. Our contribution is the definition of what we call a 'relatively monoidal functor'. The main example will be the distribution functor $\mathcal{D}$.

We start from more general assumptions. Let $F\colon \mathbf{A} \to \mathbf{B}$ be a functor between categories $\mathbf{A}, \mathbf{B}$ with finite limits, and let $A \in \mathbf{A}$ be a fixed object. We form the slice category $\mathbf{B}/F(A)$ and the functor $F_A\colon \mathbf{A} \to \mathbf{B}/F(A)$ by:

$$F_A(X) = \begin{pmatrix} F(X \times A) \\ \downarrow{\scriptstyle F(\pi_2)} \\ F(A) \end{pmatrix} \quad \text{and } F_A(f) = F(f \times \mathrm{id}_A).$$

We recall that the functor F is *monoidal* if there are maps $\zeta\colon 1 \to F(1)$ and $\xi\colon F(X) \times F(Y) \to F(X \times Y)$, where ξ is a natural transformation in a diagram as on the left below.

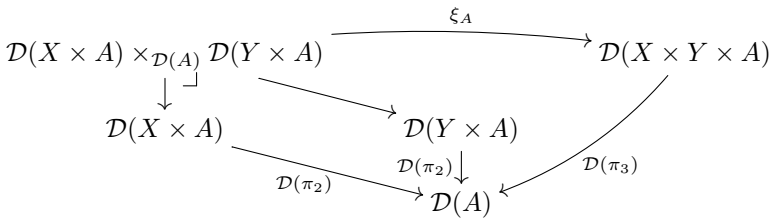
These ζ, ξ should interact appropriately, in 'unit' and 'associativity' laws, see [2] for details. The *merge* map from Section 2, in binary form, together with the unit $\eta\colon 1 \to \mathcal{D}(1)$ makes the distribution functor $\mathcal{D}$ monoidal, as depicted above on the right.

In a slice category, like $\mathbf{B}/F(A)$, products are given by pullbacks. The terminal object is the identity map $\mathrm{id}_{F(A)}$ on $F(A)$.

▶ **Definition 6.** A functor $F\colon \mathbf{A} \to \mathbf{B}$ is called *relatively monoidal* if for each object $A \in \mathbf{A}$ the functor $F_A\colon \mathbf{A} \to \mathbf{B}/F(A)$ is monoidal in the ordinary sense, say via maps $\zeta_A\colon \mathrm{id}_{F(A)} \to F_A(1)$ and $\xi_A\colon F_A(X) \times_{F(A)} F_A(Y) \to F_A(X \times Y)$, where $\times_{F(A)}$ is the product in $\mathbf{B}/F(A)$, that is the pullback over $F(A)$ in $\mathbf{B}$.

There is a canonical choice for ζ_A, namely the map $F(\langle !, \mathrm{id}\rangle)\colon F(A) \xrightarrow{\cong} F(1 \times A)$. If the functor F satisfies $F(1) \cong 1$, then $\mathbf{B}/F(1) \cong \mathbf{B}$; in that case relatively monoidal implies monoidal. There is more to say when F is a monad instead of a functor, but we shall not expand at this stage. Instead, we turn to our motivating example.

▶ **Example 7.** This definition generalises what we have used in Subsection 4.1 for the distribution functor $\mathcal{D}$ on **Sets**. We can define a map ξ_A in the slice category $\mathbf{Sets}/\mathcal{D}(A)$ in the following diagram.



For distributions $\varphi \in \mathcal{D}(X \times A)$ and $\psi \in \mathcal{D}(Y \times A)$ in the pullback on the left — so that $\mathcal{D}(\pi_2)(\varphi) = \lambda a.\, \sum_x \varphi(x, a) = \lambda a.\, \sum_y \psi(y, a) = \mathcal{D}(\pi_2)(\psi)$ — we define $\xi_A(\varphi, \psi) \in \mathcal{D}(X \times Y \times A)$ like in (14) as:

$$\xi_A(\varphi, \psi)(x, y, a) = \frac{\varphi(x, a) \cdot \psi(y, a)}{\sum_x \varphi(x, a)} = \frac{\varphi(x, a) \cdot \psi(y, a)}{\sum_y \psi(y, a)}.$$

It is not hard to see that $\mathcal{D}(\pi_3)(\xi_A(\varphi, \psi)) = \mathcal{D}(\pi_2)(\varphi) = \mathcal{D}(\pi_2)(\psi)$, so that ξ_A is a well-defined operation in the slice category $\mathbf{Sets}/\mathcal{D}(A)$. Moreover, together with the map $\zeta_A = \mathcal{D}(\langle !, \mathrm{id}\rangle)\colon \mathrm{id} \xrightarrow{\cong} \mathcal{D}_A(1)$ in $\mathbf{Sets}/\mathcal{D}(A)$ it makes the functor $\mathcal{D}_A\colon \mathbf{Sets} \to \mathbf{Sets}/\mathcal{D}(A)$ monoidal. Hence $\mathcal{D}$ is relatively monoidal.

7 Conclusions

This paper is a first step towards a bridge between the areas of Bayesian networks and coalgebra, making systematic use of (discrete probability) distributions, fuzzy predicates, probabilistic validity and conditional distributions. Such a bridge may bring more mathematical rigour to the field of Bayesian networks, and lead to new notions and ideas in the field of coalgebra and category theory, like in Section 6. Obviously, this paper only scratches the surface and many more issues remain, for instance about d-separation, bisimulation, continuous probability, quantum networks, or causality. But hopefully it leads to a fruitful exchange of ideas.

References

1 S. Awodey. *Category Theory*. Oxford Logic Guides. Oxford Univ. Press, 2006.

2 S. Eilenberg and M. Kelly. Closed categories. In S. Eilenberg, D. Harrison, S. MacLane, and H. Röhrl, editors, *Proc. Conf. on Categorical Algebra. LaJolla 1965*, pages 421–562. Springer, Berlin, 1966.

3 B. Fong. Causal theories: A categorical perspective on Bayesian networks. Master's thesis, Univ. of Oxford, 2012. see `http://arxiv.org/abs/1301.6201`.

4 R. Furber and B. Jacobs. Towards a categorical account of conditional probability, 2013. QPL 2013, see arxiv.org/abs/1306.0831.

5 J. Henson, R. Lal, and M. Pusey. General probabilistic theories on arbitrary causal structures, 2014. QPL 2014, see `arxiv.org/abs/1308.4557`.

6 T. Leinster. *Basic Category Theory*. Cambridge Studies in Advanced Mathematics. Cambridge Univ. Press, 2014.

7 S. Mac Lane. *Categories for the Working Mathematician*. Springer, Berlin, 1971.

8 L. Moss. Coalgebraic logic. *Ann. Pure & Appl. Logic*, 96(1-3):277–317, 1999. *Erratum* in *Ann. Pure & Appl. Logic*, 99(1-3):241–259, 1999.

9 J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Graduate Texts in Mathematics 118. Morgan Kaufmann, 1988.

10 S. Russel and P. Norvig. *Artificial Intelligence. A Modern Approach*. Prentice Hall, 2003.

11 E. de Vink and J. Rutten. Bisimulation for probabilistic transition systems: a coalgebraic approach. *Theor. Comp. Sci.*, 221:271–293, 1999.