

FOUNDATIONS FOR BAYESIAN NETWORKS

Jon Williamson*

in ‘Foundations of Bayesianism’
eds. David Corfield & Jon Williamson
Kluwer Applied Logic Series 2001

Bayesian networks are normally given one of two types of foundations: they are either treated purely formally as an abstract way of representing probability functions, or they are interpreted, with some causal interpretation given to the graph in a network and some standard interpretation of probability given to the probabilities specified in the network. In this chapter I argue that current foundations are problematic, and put forward new foundations which involve aspects of both the interpreted and the formal approaches.

One standard approach is to interpret a Bayesian network objectively: the graph in a Bayesian network represents causality in the world and the specified probabilities are objective, empirical probabilities. Such an interpretation founders when the Bayesian network independence assumption (often called the causal Markov condition) fails to hold. In §2 I catalogue the occasions when the independence assumption fails, and show that such failures are pervasive. Next, in §3, I show that even where the independence assumption does hold objectively, an agent’s causal knowledge is unlikely to satisfy the assumption with respect to her subjective probabilities, and that slight differences between an agent’s subjective Bayesian network and an objective Bayesian network can lead to large differences between probability distributions determined by these networks.

To overcome these difficulties I put forward logical Bayesian foundations in §5. I show that if the graph and probability specification in a Bayesian network are thought of as an agent’s background knowledge, then the agent is most rational if she adopts the probability distribution determined by the

*Department of Philosophy, King’s College, London. jon.williamson@kcl.ac.uk.

Bayesian network as her belief function. Specifically, I argue that causal knowledge constrains rational belief via what I call the causal irrelevance condition, and I show that the distribution determined by the Bayesian network maximises entropy given the causal and probabilistic knowledge in the Bayesian network.

Now even though the distribution determined by the Bayesian network may be most rational from a logical point of view, it may not be close enough to objective probability for practical purposes. I show in §6 that by adding arrows to the Bayesian network according to a conditional mutual information arrow weighting, one can decrease the cross entropy distance between the Bayesian network distribution and the objective distribution. This can be done within the context of constraints on the Bayesian network which limit its size and the time taken to calculate probabilities from the network, in order to minimise computational complexity.

This leads to two-stage foundations for Bayesian networks:^{§4} first adopt the probability function determined by a Bayesian network (this, according to the logical Bayesian interpretation, is the best subjective probability function one can adopt given the knowledge encapsulated in the network), and secondly refine the Bayesian network to better fit objective probability (this process of calibration is required by empirical Bayesianism).¹

To start with I shall give an introduction to Bayesian networks and their foundations in §1, before proceeding to criticisms of the standard interpretations of Bayesian networks in §2 and §3. The remainder of the paper will be taken up with my suggestions for new foundations.

§1

BAYESIAN NETWORKS

Suppose we have a domain of N variables, C_1, \dots, C_N , each of which takes finitely many values, $v_i^1, \dots, v_i^{K_i}$, $i = 1, \dots, N$. A *literal* is an expression c_i of the form $C_i = v_i^j$ and a *state* is a conjunction of literals. A *Bayesian network* consists of a directed acyclic graph, or *dag*, G over the nodes C_1, \dots, C_N

¹See the introduction to this volume for more on the distinction between logical and empirical Bayesianism. Such forms of Bayesianism are often referred to as ‘objective’ Bayesian positions, and confusion can arise because physical or empirical probability (frequency, propensity or chance) is often called ‘objective’ probability in order to distinguish it from Bayesian ‘subjective’ probability. In this chapter I will draw the latter distinction, using ‘objective’ to refer to empirical interpretations of causality and probability that are to do with objects external to an agent, and using ‘subjective’ to refer to interpretations of causality and probability that depend on the perspective of an agent subject.

together with a set of specifying probability values $S = \{p(c_i|d_i) : c_i \text{ is a literal involving node } C_i \text{ and } d_i \text{ is a state of the parents of } C_i \text{ in } G, i = 1, \dots, N\}$.² Now, under an independence assumption,³ namely that given its parents D_i , each node C_i is probabilistically independent of any set S of other nodes not containing the descendants of C_i , $p(c_i|d_i \wedge s) = p(c_i|d_i)$, a Bayesian network suffices to determine a joint probability distribution p over the nodes C_1, \dots, C_N .⁴ Furthermore, any probability distribution on C_1, \dots, C_N can be represented by some Bayesian network.

Bayesian networks are important in many areas where probabilistic inference must be performed efficiently, such as in expert systems for artificial intelligence. Diagnosis constitutes a typical problem area for expert systems: here one is presented with a state of symptoms s and, under the probabilistic approach to diagnosis, one must find $p(c_i|s)$ for a range of causal literals c_i .⁵ Depending on the structure of the graph G , both the number of specifiers required to determine a probability distribution p and the computational time required to calculate $p(c_i|s)$ may be substantially lower for a Bayesian network under the independence assumption than for a representation of p which makes no assumptions. Thus Bayesian networks can offer key pragmatic advantages over formalisms without an assumption like independence.

There are two main types of philosophical foundations given to Bayesian networks. One can treat Bayesian networks as *abstract structures*, and use machine learning techniques to learn from a database of past case data (for instance of the symptoms and diagnoses of past patients) a Bayesian network that represents, or represents an approximation to, a target probability distribution.⁶ More commonly, Bayesian networks are *interpreted*. Here the graph is taken to represent a causal structure, either objective or subjective. In the former case the graph contains an arrow from C_i to C_j if C_i is a direct cause of C_j , but in the subjective case the graph represents the causal knowledge of an agent X , with an arrow from C_i to C_j if X believes, or knows, that C_i is a direct cause of C_j . The specified probabilities are also given an interpretation, either objective in terms of empirical frequencies, propensities or chances, or more often subjective in terms of degrees of ra-

²If C_i has no parents, $p(c_i|d_i)$ is just $p(c_i)$.

³The Bayesian network independence assumption is often called the Markov or causal Markov condition.

⁴The joint distribution p can be determined by the *direct method*: $p(c_1 \wedge \dots \wedge c_N) = \prod_{i=1}^N p(c_i|d_i)$ where d_i is the state of the direct causes of C_i which is consistent with $c_1 \wedge \dots \wedge c_N$. Alternatively p may be determined by potentially more efficient *propagation algorithms*. See [Pearl 1988] or [Neapolitan 1990] here and for more on the formal properties of Bayesian networks.

⁵See [Williamson 2000] for more on the probabilistic approach to diagnosis.

⁶See [Jordan 1998].

tional belief. Finally the independence assumption is posited as a relation between the causal interpretation and the interpretation of probability.

In my view the most important limitation of the abstract approach is that there is often not enough initial data for it to get off the ground. The abstract approach requires a database of past case data, but there may simply not be enough such data to invoke a machine learning algorithm for generating a Bayesian network. Furthermore, new case data may trickle in slowly and it may take a while before the learning algorithm yields dependable results. Even if there is plenty of data, the data may not be reliable enough to generate a reliable network — in my experience this is a significant problem, since different people often measure or categorise variables in different ways even when collecting data for the same database. There is also a difficulty when certain variables are not measured at all: diagnostic data, for example, rarely includes the presence or absence of every possible symptom of a patient, but just the most significant symptoms, and the symptoms considered most significant are subject to biases of individual doctors. In sum, the abstract approach is not appropriate for applications which require an expert system operating right from the outset, but where the data is not available, is of poor quality, or is subject to mixtures of unknown biases. However the interpreted approach does not face this sort of problem: an expert can often from the outset provide qualitative causal knowledge, subjective degrees of belief and even estimates of objective probabilities, and this information can be used to construct a Bayesian network right away — no past case data is required.

On the other hand the interpreted approach also has its problems, largely to do with the status of the independence assumption.⁷ In the next two sections I shall outline these problems with the independence assumption and then go on to develop a hybrid methodology incorporating aspects of both the interpreted and abstract accounts: the basic idea behind the hybrid methodology is to form an initial Bayesian network from expert knowledge, and to further refine this network in the light of new case data. First we shall tackle the problems with an objective interpretation, and then investigate the subjective approach in §3.

⁷One problem that I will not consider here is the *knowledge elicitation problem*: the expert may find it hard to articulate her knowledge, and the elicitation process can be quite slow.

§2

OBJECTIVE NETWORKS

Under an objective interpretation, the Bayesian network independence assumption makes a substantive claim about the relationship between objective causality and objective, empirical probability. I will show here that this claim is highly problematic, rendering an objective interpretation inadequate.

It will be useful to note that the *principle of the common cause* is a logical consequence of the independence assumption.⁸ The principle of the common cause claims the following. Suppose two variables are probabilistically dependent and neither causes the other, then

- **existence:** they have one or more causes in common,⁹ and
- **screening:** they are probabilistically independent conditional on those common causes.

We can exploit the link between independence and the common cause principle because when an objective interpretation is given to both principles one can find many counterexamples to the latter principle which thereby contradict the former. In effect we can translate doubts about probabilistic analyses of causality in the philosophical literature — such analyses often appeal to the objectively-interpreted principle of the common cause — into doubts about the objective interpretation of Bayesian networks. Many of the counterexamples are well-known and, when considered in isolation, thought to be so unusual as to be unimportant, or thought to be susceptible to particular rebuttals. I want to provide a taxonomy of the counterexamples in order to show that the problem is more widespread than often considered

⁸This principle is due to Reichenbach (see [Reichenbach 1956], §19, pages 157-167). It is also often assumed as a basis for statistical experimentation — [Fisher 1935]. One can see that the principle of the common cause is a consequence of the independence assumption by generalising the following example in the obvious way. Suppose we have a Bayesian network with graph $A \rightarrow B, C \rightarrow D$. Thus neither B nor D cause the other, nor do they have a common cause. B and D must then be unconditionally probabilistically independent since for literals b and d on B and D respectively, their joint probability $p(b \wedge d) = \sum_{a,c} p(b|a)p(a)p(d|c)p(c) = [\sum_a p(b|a)p(a)][\sum_c p(d|c)p(c)] = p(b)p(d)$, where the first equality follows from the direct decomposition of probability in a Bayesian network (see [Neapolitan 1990] theorem 5.1 for example).

⁹Existence of a common cause resembles Mill's Fifth Canon of Inductive Reasoning: 'Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation.' [Mill 1843], page 287.

and so general that the rebuttals are either too particular or unappealing when generalised.¹⁰

I shall argue against the independence assumption by documenting two types of counterexample to the principle of the common cause: the causal variables C_i and C_j may be *accidentally correlated*, or there may be some *extra-causal constraint* which ensures that they are probabilistically correlated.¹¹ There may either be no suitable common cause to account for a correlation, contradicting the existence condition above, or if there are common causes, they will not account for all of the correlation, contradicting the screening condition.

2.1 ACCIDENTAL CORRELATIONS

Christmas trees tend to be sold when most oranges ripen and are sold. Let C represent the number of Christmas trees sold on any day and O represent the number of oranges sold on any day (C and O are random variables). Then $p(C > x | O > y) > p(C > x)$ for some suitable constants x and y . Now it seems clear that sales of Christmas trees do not cause sales of oranges, nor vice versa. Hence, some common cause must be found to explain their probabilistic dependence if the independence assumption is to hold. If there is a common cause it would have to be something like the time of year or the season. However, intuitively one does not endow the time of the year with causal powers, and there are no obvious mechanisms at play underlying any such causation. Intuitively there is no common causal explanation for the correlation — it is accidental. If such intuitions are right, then the independence assumption must fail for this causal scenario.

In order to save the independence assumption one may well be tempted to maintain that the time of year really is the common cause here. I shall call this strategy *causal extension*. The idea is that one tries to extend the intuitive concept of cause by counting intuitively non-causal variables, like the time of the year, as causal. In the context of Bayesian networks, causal extension often takes the form of an assumption that there is a ‘hidden’, ‘latent’ or ‘unmeasured’ common cause whenever two variables are found to

¹⁰A large literature touches on the independence assumption in one way or another. Thus there are criticisms (for example [Humphreys & Freedman 1996], [Humphreys 1997], [Lemmer 1993], [Lemmer 1996], [Lad 1999]) and defences (for example [Spirtes et al. 1997], [Hausman 1999], [Pearl 2000] §2.9.1) of the independence assumption which I will not cover here. I will however cover the criticisms I believe most telling and the most viable reactions to these criticisms.

¹¹‘Correlation’ is occasionally used to denote some kind of *linear* dependence, but I shall just use it as a synonym for ‘probabilistic dependence’ here.

be correlated, even when there is no intuitively plausible common cause.¹² Unfortunately, there are a number of difficulties with the strategy of causal extension. Firstly, extending the concept of cause creates epistemic problems. Identifying causal variables and the causal relationships between them is a hard problem. Any extension of the concept of cause is likely to make the task harder. In particular, it may be very difficult for an expert to provide a causal graph under the causal extension approach: one is asking the expert to identify variables that render the independence assumption valid, rather than to identify the causes and effects that she is used to dealing with. Furthermore, if one increases the number of nodes and arrows that must be considered in the graph of a Bayesian network then one risks the network becoming too complex for practical use. The amount of space required to store a Bayesian network and the amount of time required to calculate probabilities from the network both increase exponentially with the number of nodes in the worst case. This worst case occurs when the graph is dense — that is, there are many arrows in the graph. Thus causal extension is a dangerous tactic from an epistemic and practical point of view.

The second major problem is that by extending the concept of cause we are liable to lose qualities that are important to causality. Genuine causal variables tend to have various characteristics in common: for example one can normally view them as spacio-temporally localised events, and causes and effects tend to be related by physical mechanisms. If we allow variables which do not have these qualities then we can no longer be said to be explicating the notion of cause — the extension is ad hoc and the word ‘cause’ loses meaning, just becoming a synonym for ‘variable’ if the process is pursued indefinitely. This is clearly undesirable if we require a genuinely causal interpretation of the graph in the Bayesian network, as opposed to more abstract foundations.

Elliott Sober produced the following counterexample to the principle of the common cause:

Consider the fact that the sea level in Venice and the cost of bread in Britain have both been on the rise in the past two centuries. Both, let us suppose, have monotonically increased. Imagine that we put this data in the form of a chronological list; for each date, we list the Venetian sea level and the going price of British bread. Because both quantities have increased steadily in time, it is true that higher than average sea levels tend to be associated with higher than average bread prices. The two quantities are very strongly positively correlated.

¹²See [Binder et al. 1997] and [Pearl 2000] for example.

I take it that we do not feel driven to explain this correlation by postulating a common cause. Rather, we regard Venetian sea levels and British bread prices as both increasing for somewhat isolated endogenous reasons. Local conditions in Venice have increased the sea level and rather different local conditions in Britain have driven up the cost of bread. Here, postulating a common cause is simply not very plausible, given the rest of what we believe.¹³

Here Sober calls the existence of a common cause into question — there is a causal explanation of the correlation, but it is not an explanation involving *common causes*, so in a sense the correlation is accidental. Postulating a common cause conflicts with intuitions here. In particular there appears to be no common causal *mechanism*. We often appeal to non-probabilistic issues like mechanisms to help determine which correlations are causal and which are accidental. As Schlegel points out, ‘we reject a correlation between sun spots and economic cycles as probably spurious, because we know of no relating process, but accept a correlation between sun spots and terrestrial magnetic storms because there is a plausible physical relationship.’¹⁴

Besides causal extension, there is a separate line of response one can make to such counterexamples, that of *restriction*, whereby one restricts the application of the independence assumption so that it does not apply to awkward cases like Sober’s.¹⁵ This response can take one of two forms, *correlation restriction* or *causal restriction*. Regarding the former, some, such as Papineau and Price, claim that British bread prices and the Venetian water level do not have the right type of correlation for the principle of the common cause to be applied since their correlation can be predicted from the co-variation within each time-series¹⁶ or from determinism within each physical process.¹⁷ They thus attempt to avoid the counterexample to the common cause principle by restricting the principle itself. However, it should be noted that they pursue this strategy in the context of a defence of a probabilistic analysis of causality. Whether or not this move is successful in that context, it is no help here when thought of in terms of the Bayesian network framework, for restricting the principle of the common cause restricts the independence assumption too, and the reduction of a probability function to a Bayesian network is not possible without full-blown independence. Hence correlation

¹³[Sober 1988] 215.

¹⁴[Schlegel 1974] 10.

¹⁵Lakatos called this type of defence ‘monster-barring’.

¹⁶[Papineau 1992] 243.

¹⁷[Price 1992] 264.

restriction is not a viable move when considering Bayesian networks.

The other variety of restriction, causal restriction, is more promising. Here the strategy is to argue that the variables themselves are not of the sort to which the independence assumption applies. One may claim that the correlated variables are not causal variables, although this is rather implausible when it comes to the examples above. Alternatively one may accept that they are causal, but have not been individuated correctly for the independence assumption to apply. For example, the variables may need to be indexed by time,¹⁸ may need to be complete descriptions of their corresponding single-case events, or may need to be properties that can be repeatedly instantiated.

While it is possible that for any particular counterexample to independence there is another way of individuating the variables so that the dependency is removed, it is less clear that one rule of individuation will overcome all counterexamples. I have used examples which exhibit temporal correlation here because it is easy to see how such variables could be correlated, but any two events might exhibit accidental correlation, in which case alternative individuation will not help. The independence assumption rules out accidental correlation a priori, and such a restriction does not appear a priori to be any more plausible applied to one individuation than another. Thus an appeal to individuation is by no means guaranteed to overcome the problem of accidental correlation.

Causal restriction also induces epistemic problems of its own. If individuation matters then one has to do a certain amount of analysis before tackling a problem, making the application of Bayesian networks harder. Furthermore, in a particular problem one may be interested in variables which must be individuated in a way for which independence does not hold, in which case the machinery of Bayesian networks cannot be applied at all.

I have illustrated the problem of accidental correlations and introduced strategies for defending the independence assumption, including causal extension and causal restriction. These strategies are somewhat less than effective at dealing with the problem, and if they can be made to work will only do so at an epistemic and intuitive cost. In §2.2 we will see how these strategies can be applied to other common types of counterexample. Our conclusions will be much the same. Yet these costs are not ones we have to reluctantly accept. In the foundations I propose later, we will stick with our intuitive notion of cause and the individuation of variables will not matter.

¹⁸See [Spirtes et al. 1993] page 63 for example.

2.2 EXTRA-CAUSAL CONSTRAINTS

I shall now consider counterexamples to the principle of the common cause where probabilistic dependencies have an explanation that relates the dependent variables — thus the dependencies are not accidental — but where the explanation is not causal. There are a number of non-causal correlators: two causal variables can be correlated

- in virtue of their meaning,
- because they are logically related,
- because they are mathematically related,
- because they are related by (non-causal) physical laws, or
- because they are constrained by local laws or boundary conditions.

Let us look at each of these situations in turn.

First, the meanings of expressions can constrain their probabilities. 'Flu and orthomyxoviridae infection are probabilistically dependent, not because they have a common cause, but because 'flu is an example of orthomyxoviridae infection — the variables have overlapping meaning.

In response one can advocate a kind of causal restriction. One can argue that causes should be individuated so as to avoid overlapping meaning, and that one should remove a node from a Bayesian network if there is another with related meaning. But this is not always a sensible move for a number of reasons. One can lose valuable information from a Bayesian network by deleting a node, since both the original nodes may be important to the application of the network. Meaning might be related through vagueness rather than classification overlap, for example if one symptom is a patient's report of fever and another is a thermometer reading, and it may be useful to consider all such related nodes. In some cases one may even want to include synonyms in a Bayesian network, for example in a network for natural language reasoning. Furthermore, removing a node can invalidate the independence assumption if the removed node is a common cause of other nodes. Or one simply may not know that two nodes have related meaning: Yersin's discovery that the black death coincides with *Pasteurella pestis* was a genuine example of scientific inference, not the sort of thing one can do at one's desk while building an expert system.

Causal extension is no better a ploy here. One could suggest that a common cause variable called 'synonymy' or 'meaning overlap' should be introduced. But this will not in general screen off such dependencies, and as

before we have epistemic cost in terms of identifying dependencies in virtue of meaning and the likely added complexity of incorporating new variables and arrows, as well as a commitment to a counterintuitive concept of cause.

Probabilistic correlations can also be explained by logical relations. For instance, logically equivalent sentences are necessarily perfectly correlated,¹⁹ and if one sentence c logically implies sentence d , the probability of d must be greater than or equal to that of c . Thus one should be wary of Bayesian networks which involve logically complex variables. Suppose C causes complaints D , E and F , and that we have three clinical tests, one of which can determine whether or not a patient has both D and E , another tells us whether or not the patient has one of E and F , and the third tells us whether the patient has C . Thus there is no direct way of determining $p(d|c)$, $p(e|c)$ or $p(f|c)$ for literals c , d , e and f of C , D , E , and F respectively, but one can find $p(d \wedge e|c)$ and $p(e \vee f|c)$. One might then be tempted (in the spirit of causal extension) to incorporate $C \longrightarrow (D \wedge E)$, $C \longrightarrow (E \vee F)$ in one's causal graph, so that the probability specification of the corresponding Bayesian network can be determined objectively. In such a situation, however, C will not screen node $D \wedge E$ off from node $E \vee F$ and the independence assumption is not satisfied.

This problem seriously affects situations where causal relata are genuinely logically complex, as happens with context-specific causality. A may cause B *only if* the patient has genetic characteristic C : if the patient has any other genetic characteristic then there is no possible causal mechanism from A to B . Then the conjunction $A \wedge C$ is the cause of B , not A or C on their own. However, A may be able to cause D in everyone, so the causal graph would need to contain a node $A \wedge C$ and a second node A . One would not expect these two nodes to be screened off by any common causes.

Next we turn to mathematical relations as a probabilistic correlator. By way of example, consider the application of Bayesian network theory to colon endoscopy as documented in [Sucar et al. 1993] and [Kwoh & Gillies 1996]. The object is to guide the endoscope inside the colon towards the lumen, avoiding the diverticulum. A Bayesian network was used to identify the lumen and diverticulum from the endoscope image. The presence of the lumen causes a large dark region to appear on the endoscope screen while the diverticulum causes a small dark region. The size of the region can be directly measured, but its darkness was measured by its mean intensity level together with its intensity variance in the region. A Bayesian network was constructed incorporating these variables and the independence assumption was tested and found to fail: the mean and variance variables were found to

¹⁹At least according to standard axiomatisations of probability.

be correlated when, according to the causal graph under the independence assumption, they should not have been. The problem was that there is no obvious common cause for this correlation: mean and variance are related mathematically, not causally. We have that $VarX = EX^2 - (EX)^2$, where $VarX$ is the variance of random variable X , and E signifies expectation so that EX is the mean of X . To take the simplest example, if X is a Bernoulli random variable and $EX = x$ then $VarX = x(1 - x)$, making the mean and variance perfectly correlated. In the endoscopy case, the light intensity will have a more complicated distribution, but the mean value will still constrain the variance, making the mean and variance probabilistically dependent. To try to resolve this failure of the independence assumption, at first one of the two correlated nodes was removed (causal restriction). This gave some improvement in performance but suffered from significant loss of information. Next (causal extension) [Kwoh & Gillies 1996] attempted to introduce an extra common cause to screen off the correlation, but while this move improved the success rate of the Bayesian network, it raised fundamental problems. Firstly it is not clear what the new node represents (it was just called a ‘hidden node’), so a causal interpretation may no longer be appropriate for the graph. Secondly, the distribution specifying probabilities relating the new node to the other nodes had to be ascertained: this could only be done mathematically, by finding what the probabilities should be if the introduction of the new node allowed the unwanted correlation to be fully screened off, and could not be tested empirically or equated with any objective probability distribution. Therefore the Bayesian network lost both the objective causal and the objective probabilistic components of its interpretation. An objective interpretation is just not feasible, given extra-causal dependencies like this.

That extra-causal constraints include physical laws has been exemplified by Arntzenius:²⁰

Suppose that a particle decays into 2 parts, that conservation of total momentum obtains, and that it is not determined by the prior state of the particle what the momentum of each part will be after the decay. By conservation, the momentum of one part will be determined by the momentum of the other part. By indeterminism, the prior state of the particle will not determine what the momenta of each part will be after the decay. Thus there is no prior screener off.

The principle of the common cause fails here because there is nothing obvious that we can call a common cause — the existence component of the

²⁰[Arntzenius 1992] pages 227-228, from [van Fraassen 1980] page 29.

principle fails. But even if some weird and wonderful common cause could be found in such quantum situations, independence would still fail because screening condition would fail. Suppose we consider the spins B and C of two particles: B and C have values *up* or *down*. The two particles are fired such that one has spin up (represented by literal b) if and only if the other does (c). Suppose also that either one being spin up is as likely as not, $p(b) = p(c) = 1/2$, but that a common cause A is found which explains the spins, so $A \longrightarrow B, A \longrightarrow C$, and $p(b|a), p(c|a) = x > 1/2$. But since $p(b|c) = 1$, screening off is satisfied if and only if $1 = p(b|a \wedge c) = p(b|a)$, so the cause must be deterministic, a wildly inappropriate assumption in the quantum world. Thus we must conclude that there are quantum constraints on objective probability which are extra-causal.²¹

The philosophical literature also contains several examples of how local non-causal constraints and initial conditions can account for dependencies amongst causal variables. Cartwright, for instance, points out that

independence is not always an appropriate assumption to make. ... A typical case occurs when a cause operates subject to constraint, so that its operation to produce one effect is not independent of its operation to produce another. For example, an individual has \$10 to spend on groceries, to be divided between meat and vegetables. The amount that he spends on meat may be a purely probabilistic consequence of his state on entering the supermarket; so too may be the amount spent on vegetables. But the two effects are not produced independently. The cause operates to produce an expenditure of n dollars on meat if and only if it operates to produce an expenditure of $10 - n$ dollars on vegetables. Other constraints may impose different degrees of correlation.²²

Salmon²³ gives another counterexample to the screening condition. Pool balls are set up such that the black is pocketed (B) if and only if the white is (W), and a beginner is about to play who is just as likely as not to pot the black if she attempts the shot (S), and is very unlikely to pot the white otherwise. Thus if we let b, w and s be literals representing the occurrence of

²¹Note that [Butterfield 1992] looks at Bell's theorem and concludes (page 41) that, 'the violation of the Bell inequality teaches us a lesson, ... namely, some pairs of events are not screened off by their common past.' [Arntzenius 1992] has other examples and also argues on a different front against the principle of the common cause assuming determinism. See also [Healey 1991] and [Savitt 1996] pages 357-360 for a survey.

²²[Cartwright 1989] 113-114.

²³[Salmon 1980] pp. 150-151, [Salmon 1984] pp. 168-169.

B , W and S respectively, $p(b \leftrightarrow w) = 1$ and $p(b|s) = 1/2$, so $1/2 = p(w|s) \neq p(w|s \wedge b) = 1$ and the cause S does not screen off its effects B and W from each other. As Salmon says:

It may be objected, of course, that we are not entitled to infer ... that there is no event prior to B which does the screening. In fact, there is such an event — namely, the compound event which consists of the state of motion of the cue-ball shortly after they collide. The need to resort to such artificial compound events does suggest a weakness in the theory, however, for the causal relations among S , B and W seem to embody the salient features of the situation. An adequate theory of probabilistic causality should, it seems to me, be able to handle the situation in terms of the relations among these events, without having to appeal to such *ad hoc* constructions.²⁴

I would echo this sentiment in the current context: in my view an adequate objective causal-probabilistic interpretation of Bayesian networks should not have to appeal to *ad hoc* constructions. Spirtes, Glymour and Scheines give a causal-restriction defence against Salmon's counterexample by arguing that the collision should be more specifically individuated (in particular the momentum of the cue ball should be described).²⁵ Again this is less than satisfactory in the absence of a general theory as to how causes should be individuated.

A further example: repeatedly pull one of two beads (a blue bead B and red bead R , otherwise identical) out of a bag. Then $p(b|r) = 0 < 1/2 = p(b)$. But rather than saying that pulling out the red bead is a preventative of pulling out the blue bead, the correlation is explained by the set-up of the repeatable experiment: only one bead is pulled out of the bag in any trial. Here the set-up constrains the probabilities and isn't the sort of thing that counts as a cause.

In response to the problem of extra-causal constraints, one might admit defeat in problems such as the diagnosis of apparatus for the investigation of quantum mechanical systems,²⁶ or troubleshooting pool players, but maintain that most applications of intelligent reasoning may be unaffected. But extra-causal constraints occur just about anywhere, including central diagnosis problems for example. When diagnosing circuit boards, one may be constrained by the fact that two components cannot fail simultaneously

²⁴[Salmon 1980] 151 (my notation).

²⁵[Spirtes et al. 1993] 63.

²⁶As [Spirtes et al. 1993] do, pages 63-64.

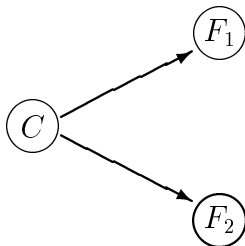


Figure 1: Failure of circuitry components.

$(F_1 \wedge F_2)$, for if one of them fails the circuit breaks and the other one cannot fail. Suppose there is a common cause C for the failures as in Figure 1. Then C fails to screen F_1 off from F_2 for $p(f_2|c \wedge f_1) = 0 \neq p(f_2|c)$. In medicine the opposite is the case: failure of one component in the human body increases the chances of failure of another, as resources are already weakened. In both these cases the constraints are very general and not the sort of thing one would want to call causes.

But why not pursue causal extension and include these extra-causal constraints in a Bayesian network? Besides the problem of a loss of the causal interpretation, we have further difficulties. Knowledge of extra-causal constraints is often in some sense superfluous to an intelligent agent's needs. An agent performing diagnosis, for instance, needs to know about causes and effects because she has to find the probabilities of various causes given some symptoms, but she is not directly concerned with facts about meaning, experimental set-ups or physical laws. Thus if there is a requirement to keep the agent's language and causal graph small, as in the Bayesian network formalism where computational complexity is an issue, extra-causal constraints are the things to leave out. Second, it may be much harder for domain experts to provide the relevant extra-causal information than the causal information. In particular, discovering all physical laws which have correlational consequences on a domain is no mean feat. Third, even if a general constraint is identified, it is often difficult to say exactly how it should be connected to the other variables in a causal graph. Should there be an arrow between the set-up of a pool table and each possible pot, or just some? Extra-causal constraints are generally symmetric while causal relations are not. Fourthly, these constraints often vary between cases in the way that causal laws don't. If the set-up of a pool table is included in a causal graph and we are interested in predicting the next pot then, since the set-up changes as play progresses, the causal graph will also have to vary radically from shot to shot. This obviously complicates the task.

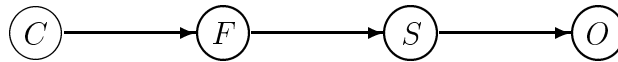


Figure 2: Christmas tree sales, festivity, spending and orange sales

Note finally that accidental and extra-causal correlations can combine to complicate matters. If two variables are accidentally correlated then a common cause is very unlikely to completely screen off that correlation. More plausibly, the common cause would account for part of the correlation, and there would be a surplus that we might call accidental. An inefficient English bakery might partly explain why the water level rises in Venice (through global warming) and also partly why bread prices rise in the UK, but the remaining bulk of the correlation might be completely accidental. Likewise direct causes of an effect may not fully screen it off from their causes. In response to our first example of accidental correlation, one might put forward some causal story: high Christmas tree sales (C) causes people to be festive (F) which causes people to spend more (S) which causes orange sales to rise (O), as in Figure 2. But even if this explains some of the correlation (and this is rather dubious), it will not explain it all, for $p(o|c) = 1$, but people spend money on many other occasions in the year and $p(o|s)$ is not much bigger than $p(o)$. So $p(o|c \wedge s) > p(o|s)$.

I hope to have shown that many types of dependency can be invoked to contest the validity of the objectively-interpreted independence assumption. Two strategies present themselves if we look for a defence against the counterexamples, causal restriction and causal extension. However each strategy is subject to epistemological, practical and intuitive difficulties, rendering an objective interpretation of Bayesian networks at worst impossible and at best undesirable.

§3

SUBJECTIVE NETWORKS

We have seen how problems arise for an objective interpretation of the components of a Bayesian network. But there is a further reason why an objective interpretation is unattractive in practice: one may simply not know of all the causal variables or causal relations relevant to a domain of interest, and one may not be able to accurately estimate the corresponding objective probabilities required in the specification of a Bayesian network. In practice our knowledge is limited, and information in a Bayesian network will often be

incomplete and inaccurate.

Thus it makes sense to relativise the Bayesian network to an agent's perspective. In this section we shall suppose that the Bayesian network expresses the knowledge of a particular agent, X say — that the graph G is interpreted as X 's representation of causality, and that the probability specification S is interpreted as containing her degrees of belief in literals conditional on parent states. The independence assumption then links the agent's picture of causality to her belief function p : if it holds then her belief function is reducible to her Bayesian network.

Does the independence assumption hold here? There is little reason to suppose that it might. X 's knowledge of causality may be very limited, and her degrees of belief may wildly differ from objective probability: according to strict-subjectivist Bayesian theory X may hold whatever beliefs she likes, as long as her belief function is formally a probability function. Yet the independence assumption is a very strong constraint, for it fixes X 's belief function given her Bayesian network, thereby restricting X 's subjectivity. If X 's causal knowledge or the degrees of belief in her probability specification were to change slightly then her other degrees of belief would have to change correspondingly, leaving no room for subjectivity with regard to these other beliefs. Therefore a strong constraint like independence does not fit well with subjectivism, whose appeal is based on the freedom it allows causal knowledge and degrees of belief.

So how can a subjective interpretation of Bayesian networks be maintained? One line of reasoning goes something like this: if independence holds objectively, and the subjective network is similar to the objective network, then the subjective distribution determined by the subjective network will be close enough to objective probability to be put to practical use. Suppose we require an expert system for diagnosis of liver disease. We may think we have a fair idea of the causal picture relating this area, and may be able to obtain estimates of the objective probabilities for a probability specification, thereby forming a Bayesian network that is in some sense close to an objective version. If the independence assumption were to hold in the objective case then one might expect it to hold approximately in the subjective case. One might further suppose that if independence approximately held in the subjective case then the probability distribution determined by the subjective network might approximate objective probability, at least closely enough for the practical purposes of liver diagnosis.

It is such a position that I want to argue against in this section. There are two flaws in the above reasoning. First, as we saw in the last section, there is often reason to doubt the independence assumption as made of objective causality and probability. Secondly, even if independence were to hold

objectively, small differences between a subjective network and the objective network can lead to significant differences in the probability distributions determined by these networks. It is this second claim that I want to argue for here.

For this argument it will be necessary to consider subjective and objective distributions and networks simultaneously, and so it will be worth spelling out the notation and concepts clearly in advance. The objective probability distribution is p^* . We also have an objective Bayesian network consisting of causal graph G^* and the associated probability specification S^* . Independence is assumed to hold of objective causality G^* with respect to objective probability p^* , and this has the repercussion that the objective network (G^*, S^*) determines p^* . Agent X has a subjective Bayesian network consisting of causal graph G and associated probability specification S . This subjective network (G, S) determines probability function p under the independence assumption. The question of whether independence holds subjectively and p matches X 's full belief function is not of concern here. Instead, we are concerned with the above alternative justification of the subjective interpretation which claims that if the subjective network (G, S) closely resembles the objective network (G^*, S^*) then the function p will be close enough to objective probability p^* to be of practical use. I argue that differences between the objective and subjective networks that are likely to occur in practice will yield significant differences between resulting probability distributions.

It will be useful to distinguish two types of difference between the subjective and objective networks: differences between the causal graphs G and G^* and differences between the probability specifications S and S^* .

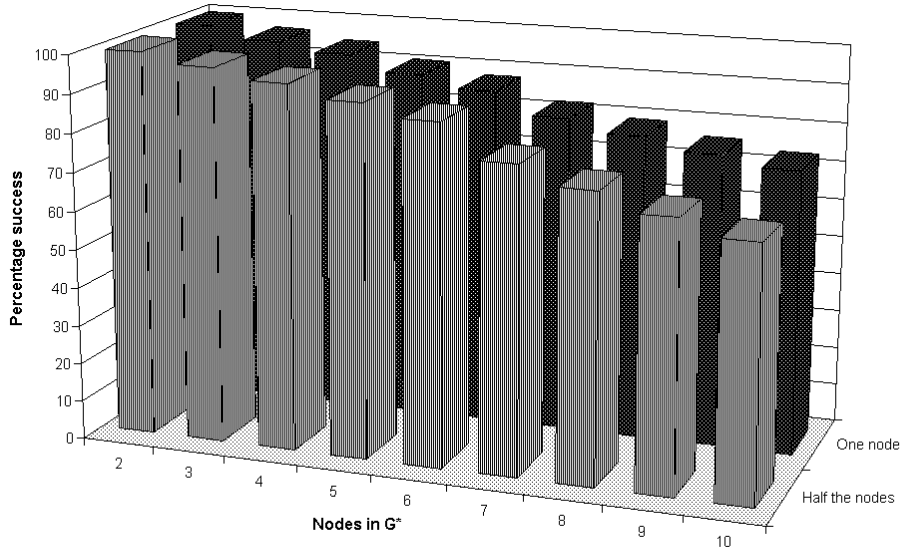
3.1 CAUSAL SUBJECTIVITY

First I shall argue as follows. Even if we make the assumption that independence holds objectively, we assume that X 's belief specification S consists of objective probabilities, and assume that her causal knowledge is correct (G is a subgraph of G^*), then if, as one would expect, her causal knowledge is incomplete (a strict subgraph), p may be not be close enough to p^* for practical purposes.

There are two basic types of incompleteness. X may well not know about all the variables (G has fewer nodes than G^*) or even if she does, she may not know about all the causal relations between the variables (G has fewer arrows than G^*).

To deal with the first case, suppose G is just G^* minus one node C and the arrows connecting it to the rest of the graph. Even if G^* satisfies

Figure 3: Nodes removed.



independence with respect to p^* then G can only be guaranteed (for all p^*) to satisfy independence if all the direct causes of C are direct causes of C 's direct effects, each pair D, E of its direct effects have an arrow between them say from D to E , and the direct causes of each such D are direct causes of E .²⁷ Needless to say, such a state of affairs is rather unlikely and a failure of independence will have practical repercussions.

I ran a simulation to indicate just how close the subjectively-determined distribution p will be to the objective distribution p^* , the results of which form Figure 3. The bars in the background of the graph show the performance of Bayesian networks formed by removing a single node and its incident arrows from networks known to satisfy independence. For $N = 2, \dots, 10$ I randomly generated Bayesian networks on N nodes, and for each net removed a random node, chose a random state of nodes s and calculated $p(c|s)$ for each literal c not in s . The new networks were deemed successful if their values for $p(c|s)$ differed from the values determined by the original network by less than 0.05, that is, $|p(c|s) - p^*(c|s)| < 0.05$. For each N the percentage success was calculated over a number of trials²⁸ and each bar in the chart represents such a percentage. The bars in the foreground of the graph represent the

²⁷See [Pearl et al. 1990] 82.

²⁸At least 2000 trials for each N , and more in cases where convergence was slow.

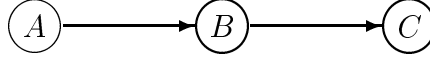


Figure 4: Objective causal graph G^* .



Figure 5: B and its incident arrows removed.

percentage success where half the nodes²⁹ and their incident arrows were removed.

Such experiments are computationally time-consuming and only practical for small values of N . While one should be wary of reading too much into a small data set, the results do suggest a trend of decreasing success rate as the size of the networks increase. Thus it appears plausible that if one removes a node and its incident arrows from a large Bayesian network that satisfies independence, then the resulting network will not be useful, in the sense that the probability values it determines will not be sufficiently close to objective probability. Moreover, removing more nodes from a Bayesian net is likely to further reduce its probability of success, as the graph shows.

This trend may be surprising, in that if one removes a node from a large causal graph one is changing a smaller portion of it than if one removes a node from a small graph, so one might expect that removing a node changes the resulting distribution less as the original number of nodes N increases. But one must bear in mind that the independence assumption is non-local: removing a node can imply an independency between two nodes which are very far apart in the graph. Thus removing a node from a small graph is likely to change fewer implied independencies than removing a node from a large graph.

Of course one may complain that such a simulation is unrealistic in some way. For instance, if one doesn't know about some intermediary cause in an objective causal graph, one may yet know about the causal chain on

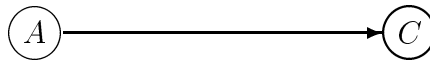
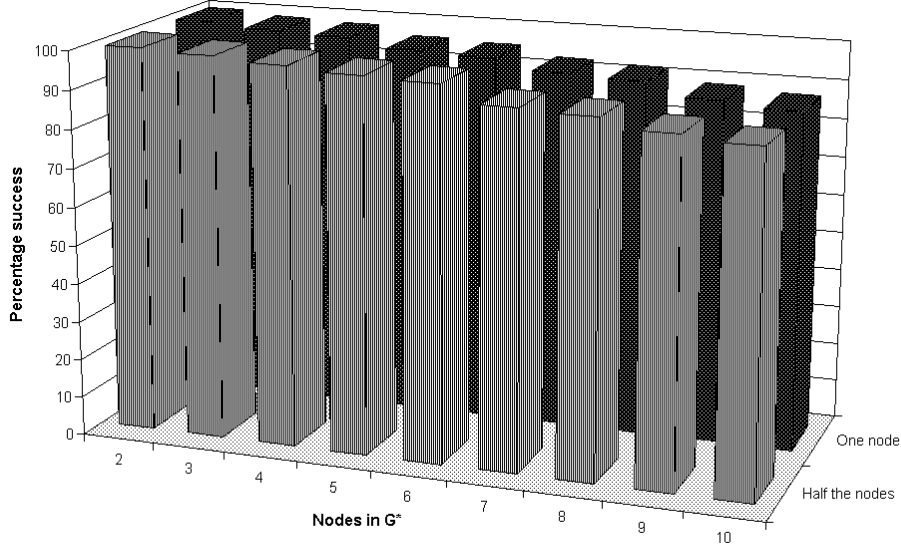


Figure 6: B removed but its incident arrows redirected.

²⁹In fact the nearest integer less than or equal to half the nodes was chosen.

Figure 7: Nodes removed - arrows re-routed.



which it exists. Thus if Figure 4 represents the objective causal graph and one doesn't know about B , one may know that A causes C , as in Figure 6 rather than Figure 5. In this case removing B 's incident arrows introduces an independence assumption which is not implied by the original graph, whereas redirecting them does not. In simulations I found that while redirecting rather than removing arrows improved success (see Figure 7) the qualitative lesson remained: the general trend was still that success decreases as the number of nodes increases.

There is another way that the simulation may be unrealistic. Some types of cause may be more likely to be unknown than others, so perhaps one should not remove a node at random in the simulation. However, if we adjust for this factor we should not expect our conclusions to be undermined. To the extent that effects are more likely to be observable and causes to be unobservable, one will be more likely to know about nodes in the latter parts of causal chains than in the earlier parts. But while removing a leaf in a graph will not introduce any new independence constraints, removing common causes can do so. Thus if X is less likely to know about causes than effects, her subjective causal graph is even less likely to satisfy independence than one with nodes removed at random.

There may be other factors which render the simulations inappropriate,

based on the way the networks are chosen at random. Here I made it as likely as not that two nodes have an arrow between them, and as likely as not that an arrow is in one direction as in another, while maintaining acyclicity. Thus the graphs are unlikely to be highly dense or highly sparse. I chose the specifying probabilities uniformly over machine reals in $[0, 1]$. Roughly half the nodes ($N/2$ nodes if N was even otherwise $(N - 1)/2$ nodes) were chosen to be symptoms in s and the nodes and their values were selected uniformly. In the face of a lack of knowledge about the large-scale structure of the objective causal graph I suggest these explications of ‘at random’ are appropriate. In any case, the trend indicated by the simulation does not seem to be sensitive to changes in the way a network is chosen at random.

In sum then, for a G^* large enough to be an objective causal graph the removal of an arbitrary node is likely to change the independencies implied by the graph, and to change the resulting distribution determined by the Bayesian network. This much is arguably true whether or not the objective situation (G^*, p^*) satisfies independence itself, for if independence fails, removing arbitrary nodes is hardly likely to make it hold.

Having looked at what happens when agent X is ignorant of causal variables, we shall now turn to the case where she is ignorant of causal relations.

Suppose then that G is formed from G^* by deleting an arrow, say from node C_i to node C_j . Then G can not be guaranteed to satisfy independence with respect to p^* . For suppose C_i, D_1, \dots, D_k are the direct causes of C_j in G^* . Then the independence of G with respect to p^* requires that C_i be independent of C_j , conditional on D_1, \dots, D_k , which is not implied by the independence of G^* with respect to p^* .

The situation is worse if the following condition holds, which I shall call the *dependence* principle.³⁰ This corresponds to the intuition that a cause will either increase the probability of an effect, or, if it is a preventative, make the effect less likely. More precisely,

- **dependence:** if C_i, D_1, \dots, D_k are the direct causes of C_j then C_i and C_j are probabilistically dependent conditional on D_1, \dots, D_k : there are some literals c_i and c_j of C_i and C_j and some state d of D_1, \dots, D_k such that $p^*(c_j|c_i \wedge d) \neq p^*(c_j|d)$, as long as these probabilities are non-extreme (that is, neither 0 nor 1).

Now if G^* satisfies dependence with respect to p^* , the arrow between C_i and C_j is removed to give G as before, and the probabilities are non-extreme, the independence assumption will *definitely fail* for G with respect

³⁰See [Williamson 1999] for a defence of this principle. Note that the dependence principle is a partial converse to the independence assumption.

to p^* . This is simply because the independence of G with respect to p^* requires that C_i and C_j be independent conditional on D_1, \dots, D_k which contradicts the assumption that dependence holds for G^* with respect to p^* . Note that this result only depends on the local situation involving C_i , C_j and the other direct causes D_1, \dots, D_k of C_j , so that further changes elsewhere in the graph cannot rectify the situation.³¹ Note also that this result does *not* require that objective causality G^* satisfy independence with respect to objective probability p^* . Thus if the dependence principle holds of causality in the world it is extremely unlikely that independence will hold of a subjective causal theory.

Of course, we are arguing against independence by appealing to an alternative principle here and the sceptical reader may not be convinced by this last argument. But we can perform simulations as before to indicate the general trends. The back row of Figure 8 represents the results of the same simulation as before (the dependence principle is not assumed to hold), except with a random arrow rather than a node removed. In this case there is no clear downward trend, but success rate is uniformly low. If more arrows are removed, then for all but small N the resulting network is less likely still to satisfy independence, as the front row of Figure 8 shows, and again we see a downward trend as the number of nodes in G^* increases.

In sum, causal subjectivity can lead to a significant difference between the subjective and objective probability distributions.

3.2 PROBABILISTIC SUBJECTIVITY

Turning now to X 's degrees of belief, it is not hard to see how p can differ from p^* . We suppose that the objective situation satisfies independence, and that X 's causal graph G matches the objective causal graph G^* . However, if her specification S differs from the objective specification then the probability function p determined by the subjective network (G, S) would not be expected to agree exactly with p^* . The back row of Figure 9 shows what happens if one of the nodes has its associated probability specifiers perturbed by 0.03, the middle row shows what happens if half the nodes' probabilities are perturbed by 0.03, and the front rows gives the case where all nodes have their probabilities perturbed.

In practice probabilistic and graphical subjectivity will occur together, making it even less likely that p is close enough to p^* for practical purposes. The back row of Figure 10 shows what happens if a node is removed (arrows

³¹If one or more of the other direct causes or their arrows to C_j are also absent in G , then independence may be reinstated, although this would be a freak occurrence and the extra change may break a further independence relation elsewhere in the graph.

Figure 8: Arrows removed.

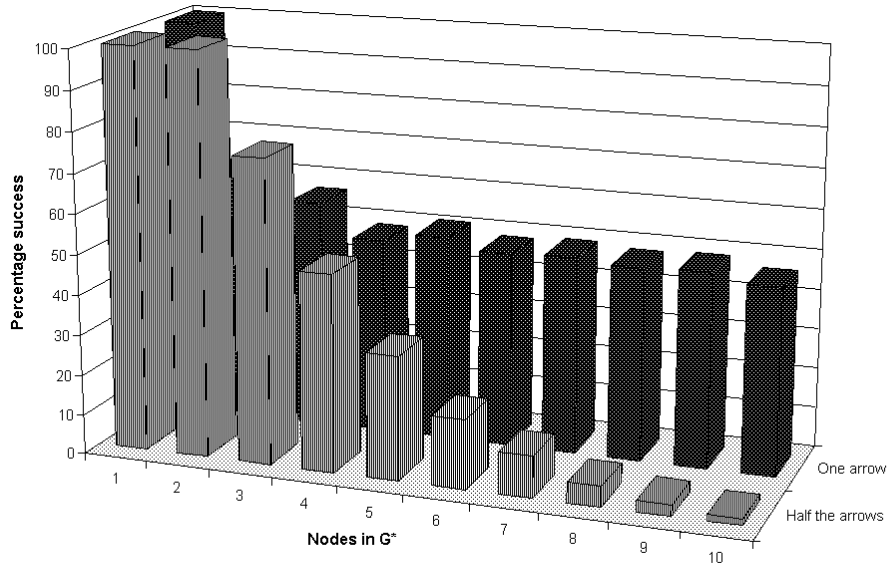


Figure 9: Node probabilities perturbed.

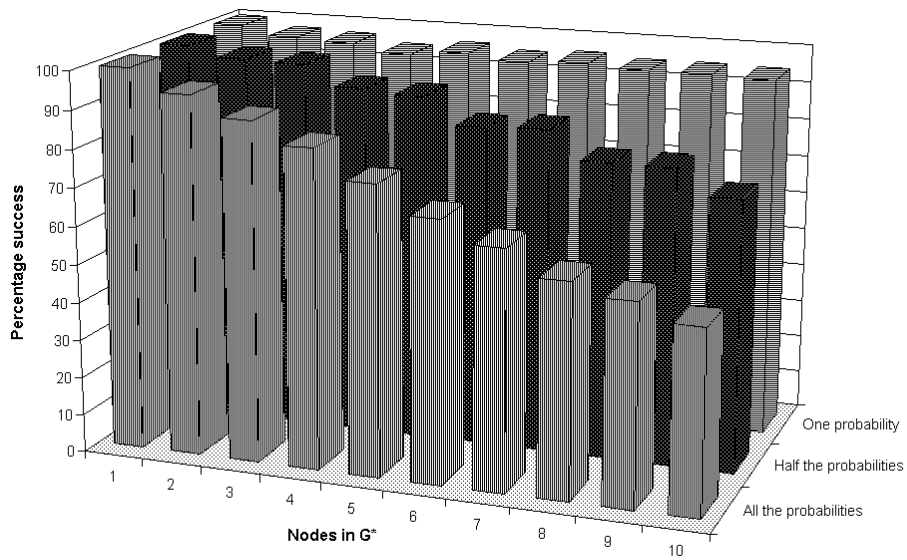
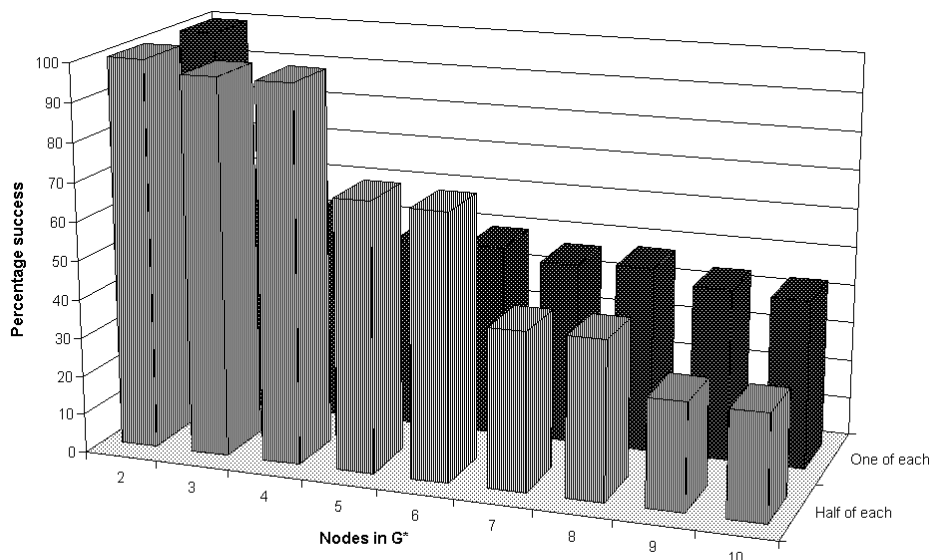


Figure 10: Nodes and arrows removed, node probabilities perturbed.



re-routed), then an arrow is removed, and then one node's probabilities are perturbed by 0.03. The front row shows what happens if half the nodes then half the remaining arrows are removed, then half the remaining nodes are perturbed.

Thus subjectivity in a Bayesian network can lead, significantly often, to practical problems: the distribution determined by a subjective network may differ too much from the objective distribution to be of practical use.

§4

TWO-STAGE BAYESIAN NETWORKS

We have seen some of the problems that face interpretations of Bayesian networks. The independence assumption can fail for an objective interpretation because correlations may be accidental or have non-causal explanations. Independence can hardly be expected to hold for a subjective interpretation — the agent's Bayesian network will generally give rise to a probability function p which differs from her true belief function — but more importantly p is also likely to differ from objective probability, which upsets the alternative justification of subjective networks.

I want to argue for another view of Bayesian networks, which I believe rests on firmer foundations. The view I put forward here initially adopts a subjective interpretation, where the graph in the Bayesian network is an agent's representation of causal structure and the probability specifiers are her degrees of rational belief. I acknowledge the fact that, according to the above arguments, the distribution specified by an agent's Bayesian network may not be close enough to the objective distribution to be of much practical use, but I argue that it is a good starting point, and can be refined to better approximate reality. This gives a *two-stage methodology* where stage one is the representation of X 's belief function p by an initial Bayesian network and stage two is the further refinement of the network. In terms of foundations, stage one yields a subjective interpretation (but a different subjective interpretation to those given in §3), while stage two borrows techniques from the abstract approach in order to deliver a network whose distribution more closely approximates the objective distribution (and in the process of refinement the causal interpretation may be dropped as we shall see).

Two key questions require attention before we can be convinced of these two-stage foundations for Bayesian networks. Firstly, how can stage one be justified? I have argued against a strict subjective interpretation, and so must somehow demonstrate that some other kind of subjective interpretation of the Bayesian network is a good starting point. I shall do this in the rest of this section and the next section. Secondly, how can stage two be performed? I shall discuss the refinement of Bayesian networks in §6.

I shall interpret X 's Bayesian network as her background knowledge: the causal graph G contains her knowledge of causal variables and their causal relations, and the probability specification S is her knowledge of conditional probabilities of causes given parent-states.³² The independence assumption may then be used to determine X 's degrees of belief from her background knowledge: her full belief function will be the probability function determined by the Bayesian network on G and S under the independence assumption.

Thus independence is no longer a substantive assumption linking the agent's causal graph with some pre-determined rational belief function, it is a *logic*, used to derive undetermined degrees of belief from those that are given in X 's probability specification.

The central issue then is how we can justify the use of the independence assumption as a means of determining a rational belief function.

This issue of finding a single rational belief function given some back-

³²I shall leave it open as to whether these probabilities are taken to be estimates of objective probabilities or informed degrees of belief. It suffices that they count as knowledge and may be used to guide X 's other beliefs.

ground knowledge has received plenty of attention in the literature. Approaches range from Laplace's *principle of indifference* to Jaynes' *maximum entropy principle*. The former says that if X is indifferent as to which of J alternatives is true then she should believe each of them to degree $1/J$. The latter explicates and generalises the former as follows. A probability function over C_1, \dots, C_N may be fully specified by specifying values for each of the parameters $x^{k_1, \dots, k_N} = p(C_1 = v_1^{k_1} \wedge \dots \wedge C_N = v_N^{k_N})$, where $v_i^{k_i} \in \{v_i^1, \dots, v_i^{K_N}\}$ for $i = 1, \dots, N$. We have the constraints that each $x^{k_1, \dots, k_N} \in [0, 1]$, and by additivity $\sum_{k_1, \dots, k_N} x^{k_1, \dots, k_N} = 1$, together with any constraints implied by background knowledge. The maximum entropy principle says that in the absence of any further information X should select a most rational belief function by choosing the x^{k_1, \dots, k_N} subject to these constraints which maximises the entropy

$$H = - \sum_{k_1, \dots, k_N} x^{k_1, \dots, k_N} \log x^{k_1, \dots, k_N}.$$

There are several convincing justifications for the maximum entropy principle. The most well-known involves Shannon's information-theoretic interpretation of entropy as a measure of uncertainty, in which case we maximise entropy subject to some background knowledge if we determine a probability function whose informativeness is as close as possible to that of just the background knowledge itself. A second justification is based on Boltzmann's work with entropy in physics, and a third involves Paris and Vencovská's demonstration that the maximum entropy solution is the only completion to satisfy various intuitively compelling desiderata, such as language invariance.³³ Grünwald gives a fourth, game-theoretic justification: maximum entropy is the (worst-case) optimal distribution for a game requiring the prediction of outcomes under a logarithmic loss function.³⁴

Where does this leave independence and stage one of our two-stage methodology? Stage one is justified because the probability function determined by the independence assumption from the Bayesian network coincides with that determined by the maximum entropy principle, as we shall now see.

³³See [Paris 1994], [Paris & Vencovská 1997], [Paris 1999] and [Paris & Vencovská 2001] for the details of these justifications.

³⁴[Grünwald 2000].

BAYESIAN NETWORKS HAVE MAXIMUM ENTROPY

The argument for the identity of the Bayesian network and maximum entropy functions requires first making the constraints imposed by the background knowledge explicit, and next showing that if we maximise entropy subject to these constraints then we get the same solution as that determined by the Bayesian network under the independence assumption.

5.1 BACKGROUND KNOWLEDGE

Agent X 's background knowledge consists of the components of a causally interpreted Bayesian network: a causal graph and the specified probabilities of literals conditional on states of their parents. We first need to formulate this knowledge in a way that can more formally be applied to the maximum entropy procedure. Regarding the probability specification, there is no problem. We can simply maximise entropy subject to the constraints that certain probabilities, namely those in the Bayesian network specification, are fixed from the outset. However, the causal graph does not provide obvious constraints — it is of qualitative form, free from notions like entropy or probability. Therefore we need some procedure for turning the causal information into a constraint on probability.

I suggest that the causal interpretation imposes the following constraint. Suppose we are presented with the components of a Bayesian network involving variables C_1, \dots, C_N and then use these to determine a single rational belief function p_1 , whether by independence, maximum entropy or some other means. Then we find out further causal information, namely that there are some new variables D_1, \dots, D_M to be added to the causal graph, and that these variables are not causes of the current C -variables C_1, \dots, C_N . Intuitively, this new information should not affect our understanding of the original problem on the C -variables. More precisely, suppose the new information takes the form of an extension of the original causal graph where the D -variables do not cause C -variables, and an extension to the probability specification incorporating new conditional probabilities of the D -variables given their parents. If we use this new Bayesian network to determine a new rational belief function p_2 over the larger domain $C_1, \dots, C_N, D_1, \dots, D_M$, then the restriction of p_2 to the C -variables should agree with p_1 , the function based just on the C -variables. I shall call this the principle of *causal irrelevance*: learning of the new variables should be irrelevant to degrees of belief on the previous domain.

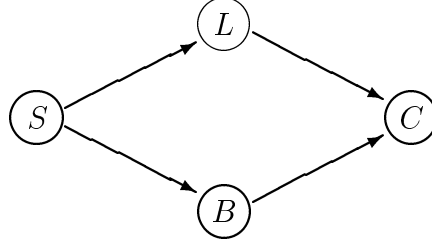


Figure 11: Smoking, lung cancer, bronchitis and chest pains.

This principle is based on an asymmetry of causation whereby information about causes can lead to information about their effects, but knowledge of effects does not provide useful information about causes. This is not to say that information about the *value* or *occurrence* of an effect is irrelevant to the question of what the value of its cause is (which is clearly wrong), but that information of the form that a variable *has an effect of unknown value* is irrelevant to its own value. The same need not be true of causes: if two variables thought to be causally unrelated are found to have a common cause, one may be wise to suppose that these variables are probabilistically dependent to a greater extent than previously thought.

Take a simple example: suppose L signifies lung cancer and B bronchitis. We know of no causal relations linking the two variables, and have the probabilities $p(l), p(b)$ for each literal l, b involving L, B respectively. We then use this information to determine a joint probability distribution p_1 over L and B . Suppose we later learn that S , smoking, is a cause of lung cancer and of bronchitis, and we find the probabilities $p(l|s), p(b|s), p(s)$ for each literal l, b, s involving L, B, S respectively. Then, because S is a common cause, we might be inclined to form a new belief function p_2 over L, B and S which renders L and B more dependent than they were under p_1 : $p_2(l|b) > p_1(l|b)$ for some literals l and b . The motivation is that if we find out b , then we now know this may be because some literal s has caused b , in which case s may also have caused l , making it more likely than we would previously have thought.

Suppose next we learn that each of lung cancer and bronchitis cause chest pains C , as in Figure 11. If we find values for $p(c|l \wedge b)$ for each literal c, l and b , and form a new belief function p_3 , the causal irrelevance condition requires that p_3 must not differ from p_2 , over S, L and B . For example, $p_3(l|b) = p_2(l|b)$, for each l and b . The idea here is that if we learn b , then knowledge of the existence of the common effect C does not give us a new way l may occur and so our degree of belief in l should not change. C is

irrelevant to S , L and B .

In sum, I shall assume that the process of determining a single rational belief function is constrained not only by the probability values in the specification of the Bayesian network, but also by the causal graph under the principle of causal irrelevance. The principle of causal irrelevance is strong enough to allow causal information to constrain rational belief, and thereby play a part in our new justification of the independence assumption, yet, unlike the independence assumption, weak enough to be uncontroversial in itself.

5.2 MAXIMISING ENTROPY

The key proposition is this:

BAYESIAN NETWORKS MAXIMISE ENTROPY

Given the probability specification and causal graph of a Bayesian network and the principle of causal irrelevance, the distribution which maximises entropy is just the distribution determined by the Bayesian network under the independence assumption.

PROOF: The strategy of the proof will be to use Lagrange multipliers to derive conditions for entropy to be maximised, and then show that the Bayesian network distribution satisfies these conditions. This straightforward method is possible for the following reason. The constraints — which consist of the specified probabilities, certain probabilities fixed by the causal graph under causal irrelevance, and additivity constraints common to all probability distributions — are linear and restrict the domain of the entropy function to a compact convex set in $[0, 1]^{K_1} \times \dots \times [0, 1]^{K_N}$,³⁵ and on that domain, entropy is a strictly concave function (as shown below). Thus the problem has a unique local maximum, the global maximum, and if the Bayesian network distribution satisfies the conditions for an optimal solution then it must be the unique global maximum.

We can see that entropy is strictly concave as follows. H is strictly concave if and only if, for any two distinct vectors a and b of the parameters x^{k_1, \dots, k_N} and $\lambda \in (0, 1)$,

$$H(\lambda a + (1 - \lambda)b) > \lambda H(a) + (1 - \lambda)H(b) \Leftrightarrow$$

$$\lambda \sum a_i \log a_i + (1 - \lambda) \sum b_i \log b_i - \sum (\lambda a_i + (1 - \lambda)b_i) \log(\lambda a_i + (1 - \lambda)b_i) > 0$$

$$\Leftrightarrow \lambda \sum a_i \log \frac{a_i}{\lambda a_i + (1 - \lambda)b_i} + (1 - \lambda) \sum b_i \log \frac{b_i}{\lambda a_i + (1 - \lambda)b_i} > 0$$

³⁵See [Paris 1994], proposition 6.1, page 66.

$$\Leftrightarrow \lambda d(a, \lambda a + (1 - \lambda)b) + (1 - \lambda)d(b, \lambda a + (1 - \lambda)b) > 0,$$

where d signifies *cross entropy*, a measure of distance of probability distributions, and a, b and $\lambda a + (1 - \lambda)b$ are non-zero since $\sum a_i = 1 = \sum b_i, \lambda \in (0, 1)$. d is well known to be non-negative and strictly positive if its arguments are distinct.³⁶ Thus $d(a, \lambda a + (1 - \lambda)b)$ is strictly positive if $a \neq \lambda a + (1 - \lambda)b$, which is true since a and b are distinct and $\lambda \in (0, 1)$. Therefore H is strictly concave and the Lagrange multiplier approach will yield the global maximum.

The next thing to do is to reformulate the optimisation problem to make it suit the Bayesian network framework. This means finding more appropriate parameters than the standard x^{k_1, \dots, k_N} mentioned above. Without loss of generality we can suppose the nodes C_1, \dots, C_N are ordered ancestrally with respect to the causal graph G in the Bayesian network: that is, all the parents of C_i in G come before C_i in the ordering.³⁷ To make the proof clearer we shall also suppose that all the probabilities in the specification are positive — we shall see later that zeros do not affect the result. Let $c_i^{k_i}$ represent the literal $C_i = v_i^{k_i}$, for $k_i = 1, \dots, K_i, i = 1, \dots, N$. The new parameters are

$$y_{i, k_i}^{k_1, \dots, k_{i-1}} = p(c_i^{k_i} | c_1^{k_1} \wedge \dots \wedge c_{i-1}^{k_{i-1}}),$$

for $i = 1, \dots, N$. The main thing to note about this parameterisation is that by the chain rule of probability,

$$x^{k_1, \dots, k_N} = \prod_{i=1}^N y_{i, k_i}^{k_1, \dots, k_{i-1}}.$$

Now we shall translate the entropy formula into this framework (in what follows we shall minimise negative entropy $-H$, which is equivalent to maximising entropy H):³⁸

$$\begin{aligned} -H &= \sum_{k_1, \dots, k_N} x^{k_1, \dots, k_N} \log x^{k_1, \dots, k_N} \\ &= \sum_{k_1, \dots, k_N} \left[\prod_{j=1}^N y_{j, k_j}^{k_1, \dots, k_{j-1}} \right] \sum_{i=1}^N \log y_{i, k_i}^{k_1, \dots, k_{i-1}} \\ &= \sum_{i=1}^N \sum_{k_1, \dots, k_N} \left[\prod_{j=1}^N y_{j, k_j}^{k_1, \dots, k_{j-1}} \right] \log y_{i, k_i}^{k_1, \dots, k_{i-1}} \end{aligned}$$

³⁶See [Paris 1994] proposition 8.5 for example.

³⁷Recall that such an ordering is always possible because of the dag structure of the causal graph.

³⁸Note that the existence and uniqueness of a maximum is independent of parameterisation.

$$= \sum_{i=1}^N \sum_{k_1, \dots, k_i} \left[\prod_{j=1}^i y_{j, k_j}^{k_1, \dots, k_{j-1}} \right] \log y_{i, k_i}^{k_1, \dots, k_{i-1}},$$

where we make this last step because for each i we can separate out

$$\sum_{k_{i+1}, \dots, k_N} \left[\prod_{j=i+1}^N y_{j, k_j}^{k_1, \dots, k_{j-1}} \right],$$

and these terms cancel to 1 by additivity of probability.

We shall deal with three types of constraints. The specification constraints are determined by those values provided in the Bayesian network specification. Causal constraints are determined by the causal graph under the causal irrelevance condition. Finally additivity constraints are imposed by the axioms of probability. While one might suspect that all these constraints would lead to a complicated optimisation problem, we will see that by adopting an inductive approach we will be able to form a Lagrangian function which only incorporates relatively few specification and additivity constraints.

Within the new framework we can write the specification constraints as

$$p(c_i^{k_i} | c_{r_1}^{k_{r_1}} \wedge \dots \wedge c_{r_L}^{k_{r_L}}) = a_{i, k_i}^{k_{r_1}, \dots, k_{r_L}},$$

where the c_{r_1}, \dots, c_{r_L} involve the parents of C_i , $r_1, \dots, r_L < i$ (thanks to the ancestral order) and $i = 1, \dots, N$.³⁹ We also have constraints imposed by additivity: $\sum_{k_i} y_{i, k_i}^{k_1, \dots, k_{i-1}} = 1$ for each k_1, \dots, k_{i-1} , $i = 1, \dots, N$.

Decomposing the entropy as $H = \sum_{i=1}^N H_i$ where

$$H_i = \sum_{k_1, \dots, k_i} \left[\prod_{j=1}^i y_{j, k_j}^{k_1, \dots, k_{j-1}} \right] \log y_{i, k_i}^{k_1, \dots, k_{i-1}},$$

we shall prove the proposition by induction on N . The case $N = 1$ is trivial since the constraints $p(c_1^{k_1}) = a_{1, k_1}$ completely determine the probability distribution over C_1 : there is nothing to do to maximise entropy and so the Bayesian network distribution, which satisfies the constraints, maximises entropy. Suppose the induction hypothesis holds for $N - 1$ and consider the case for N . It is here that we apply the principle of causal irrelevance to generate the causal constraints on the maximisation process from the causal graph. Since the variables are ordered ancestrally, the move from $N - 1$ to N essentially involves incorporating a new variable C_N which is not a

³⁹Note that the r_1, \dots, r_L depend on i . I am inclined to avoid any further subscripting however.

cause of any of the previous variables C_1, \dots, C_{N-1} . Hence if we maximise entropy on this new domain and restrict the resulting probability function to C_1, \dots, C_{N-1} then by causal irrelevance we must have maximised entropy on this smaller domain. Applying the induction hypothesis on this smaller domain $\{C_1, \dots, C_{N-1}\}$, we see that entropy is maximised if the distribution is determined by the Bayesian network on C_1, \dots, C_{N-1} . Thus for $i = 1, \dots, N-1$, the parameters $y_{i,k_i}^{k_1, \dots, k_{i-1}}$ must be fixed to $a_{i,k_i}^{k_{r_1}, \dots, k_{r_L}}$. Now H_1, \dots, H_{N-1} involve only these fixed parameters, so in order to maximise H all that remains is to maximise H_N with respect to $y_{N,k_N}^{k_1, \dots, k_{N-1}}$, subject to the specification constraints fixing the values $a_{N,k_N}^{k_{r_1}, \dots, k_{r_L}}$ and the additivity constraints $\sum_{k_N} y_{N,k_N}^{k_1, \dots, k_{N-1}} = 1$ for each k_1, \dots, k_{N-1} .

We shall now adapt the specification constraints.

Let $b^{k_{r_1}, \dots, k_{r_L}} = p(c_{r_1}^{k_{r_1}} \wedge \dots \wedge c_{r_L}^{k_{r_L}})$ and $e^{k_1, \dots, k_{N-1}} = \prod_{j < N} y_{j,k_j}^{k_1, \dots, k_{j-1}}$ be constants, fixed by having maximised entropy on C_1, \dots, C_{N-1} . Then

$$\begin{aligned} a_{N,k_N}^{k_{r_1}, \dots, k_{r_L}} b^{k_{r_1}, \dots, k_{r_L}} &= p(c_N^{k_N} \wedge c_{r_1}^{k_{r_1}} \wedge \dots \wedge c_{r_L}^{k_{r_L}}) \\ &= \sum_{k_i, i \neq r_1, \dots, r_L, N} p(c_1^{k_1} \wedge \dots \wedge c_N^{k_N}) \\ &= \sum_{k_i, i \neq r_1, \dots, r_L, N} \prod_{j \leq N} y_{j,k_j}^{k_1, \dots, k_{j-1}}, \\ &= \sum_{k_i, i \neq r_1, \dots, r_L, N} e^{k_1, \dots, k_{N-1}} y_{N,k_N}^{k_1, \dots, k_{N-1}}. \end{aligned}$$

We are now in a position to specify the Lagrangian function for the minimisation of $-H_N$:

$$\begin{aligned} \Lambda_N &= \sum_{k_1, \dots, k_N} e^{k_1, \dots, k_{N-1}} y_{N,k_N}^{k_1, \dots, k_{N-1}} \log y_{N,k_N}^{k_1, \dots, k_{N-1}} + \sum_{k_{r_1}, \dots, k_{r_L}, k_N} \lambda_{k_N}^{k_{r_1}, \dots, k_{r_L}} \times \\ &\quad \left[\sum_{k_i, i \neq r_1, \dots, r_L, N} e^{k_1, \dots, k_{N-1}} y_{N,k_N}^{k_1, \dots, k_{N-1}} - a_{N,k_N}^{k_{r_1}, \dots, k_{r_L}} b^{k_{r_1}, \dots, k_{r_L}} \right] \\ &\quad + \sum_{k_1, \dots, k_{N-1}} \mu^{k_1, \dots, k_{N-1}} \left[\sum_{k_N} y_{N,k_N}^{k_1, \dots, k_{N-1}} - 1 \right] \\ &= \sum_{k_1, \dots, k_N} \left(e^{k_1, \dots, k_{N-1}} y_{N,k_N}^{k_1, \dots, k_{N-1}} \log y_{N,k_N}^{k_1, \dots, k_{N-1}} + \right. \\ &\quad \left. \lambda_{k_N}^{k_{r_1}, \dots, k_{r_L}} \left[e^{k_1, \dots, k_{N-1}} y_{N,k_N}^{k_1, \dots, k_{N-1}} - a_{N,k_N}^{k_{r_1}, \dots, k_{r_L}} b^{k_{r_1}, \dots, k_{r_L}} \right] + \right. \end{aligned}$$

$$\mu^{k_1, \dots, k_{N-1}} \left[y_{N, k_N}^{k_1, \dots, k_{N-1}} - 1/K_N \right] \Big).$$

By Lagrange's theorem,⁴⁰ in order to find conditions for a minimum we must first check a constraint qualification. Enumerate the constraints f_1, \dots, f_J . Form a matrix A by letting each row i consist of the partial derivatives

$$\frac{\partial f_i}{\partial y_{N, k_N}^{k_1, \dots, k_{N-1}}}, 1 \leq k_j \leq K_j, j = 1, \dots, N.$$

Finally check that the rank of A is J — this is easily done and I shall avoid the details here.

Entropy is maximised if the partial derivatives of the Lagrangian are zero,

$$\frac{\partial \Lambda_N}{\partial y_{N, k_N}^{k_1, \dots, k_{N-1}}} = e^{k_1, \dots, k_{N-1}} \left[1 + \log y_{N, k_N}^{k_1, \dots, k_{N-1}} + \lambda_{k_N}^{k_{r_1}, \dots, k_{r_L}} \right] + \mu^{k_1, \dots, k_{N-1}} = 0$$

Given any such equation we can eliminate the Lagrange multiplier $\mu^{k_1, \dots, k_{N-1}}$ by finding another equation involving $k'_N \neq k_N$,

$$\frac{\partial \Lambda_N}{\partial y_{N, k'_N}^{k_1, \dots, k_{N-1}}} = 0$$

(there will always be another such equation since C_N has at least two values), and substituting to give a new equation

$$\lambda_{k_N}^{k_{r_1}, \dots, k_{r_L}} - \lambda_{k'_N}^{k_{r_1}, \dots, k_{r_L}} = \log y_{N, k'_N}^{k_1, \dots, k_{N-1}} - \log y_{N, k_N}^{k_1, \dots, k_{N-1}}$$

We next eliminate the multiplier expression on the left-hand side by finding another such equation involving k'_1, \dots, k'_{N-1} such that $k'_{r_1} = k_{r_1}, \dots, k'_{r_L} = k_{r_L}$. There will always be another such equation unless $L = N - 1$, in which case the constraints uniquely determine the Bayesian network distribution, and entropy is trivially maximised. This then gives

$$\log y_{N, k'_N}^{k_1, \dots, k_{N-1}} - \log y_{N, k_N}^{k_1, \dots, k_{N-1}} = \log y_{N, k'_N}^{k'_1, \dots, k'_{N-1}} - \log y_{N, k_N}^{k'_1, \dots, k'_{N-1}}.$$

Finally, all we need do is note that in the Bayesian network distribution the constraints are satisfied and the independence assumption implies that

$$y_{N, k_N}^{k_1, \dots, k_{N-1}} = y_{N, k_N}^{k_{r_1}, \dots, k_{r_L}} = a_{N, k_N}^{k_{r_1}, \dots, k_{r_L}},$$

$$y_{N, k_N}^{k'_1, \dots, k'_{N-1}} = y_{N, k_N}^{k'_{r_1}, \dots, k'_{r_L}} = y_{N, k_N}^{k_{r_1}, \dots, k_{r_L}} = a_{N, k_N}^{k_{r_1}, \dots, k_{r_L}},$$

⁴⁰See for example [Sundaram 1996] §5.2.1

in which case we substitute into our condition:

$$a_{N,k'_N}^{k_{r_1},\dots,k_{r_L}} - a_{N,k_N}^{k_{r_1},\dots,k_{r_L}} = a_{N,k'_N}^{k_{r_1},\dots,k_{r_L}} - a_{N,k_N}^{k_{r_1},\dots,k_{r_L}},$$

and find that it clearly holds. Thus the Bayesian network distribution is the entropy maximiser, as required.

All that remains is to point out what happens when specifiers may be zero. There are two (compatible) scenarios: if some $a_{j,k_j}^{k_{r_1},\dots,k_{r_L}} = 0$ for $j < N$ then the corresponding $e^{k_1,\dots,k_{N-1}} = \prod_{j<N} y_{j,k_j}^{k_1,\dots,k_{j-1}}$, which by the induction hypothesis is $\prod_{j<N} a_{j,k_j}^{k_{r_1},\dots,k_{r_L}}$, vanishes. This eliminates entropy terms and constraints equally, leaving fewer partial derivative conditions. These conditions are satisfied as above. The second scenario is that some $a_{N,k_N}^{k_{r_1},\dots,k_{r_L}} = 0$. In this case the Lagrangian and partial derivatives are as before, the constraints are satisfied as before, but when substituting zeros in the partial derivatives we make use of the convention, common when dealing with the cross entropy measure, that $0[\log 0 - \log 0] = 0 \log 0/0 = 0$. Thus the conditions are satisfied by null specifiers. \square

Thus we see that the independence assumption can be justified after all. The important thing to remember is that under the two-stage foundations, the independence assumption is neither a fact of causality nor even an assertion about an agent's knowledge. It is a mechanism that can be used to derive new probability statements from those in the agent's background knowledge. Independence is justified because as a logic it coincides with maximum entropy, which has well known justifications.

§6

STAGE TWO

Given background knowledge consisting of a causal graph G and associated probability specification S , we can represent the rational (maximum entropy) belief function p by the Bayesian network on G and S . This is stage one of the two-stage methodology. However, while p is rational given background knowledge, it may not bear a close enough resemblance to objective probability to be put to practical use. If that is the case then we need to transform the Bayesian network into one which more closely approximates objective probability. This is stage two of the two-stage methodology. Bayesian networks may be applied to medical diagnosis for example, or fault-finding in aeroplanes. In such high risk scenarios it is not sufficient that any decisions are deemed reasonable given a lack of relevant information: it would

be negligent not to collect enough relevant information to reliably model the objective situation.

Thus the next step is to refine the Bayesian network in the light of new information, in order to achieve greater reliability. Many of the algorithms from the extensive literature on learning Bayesian networks from data⁴¹ can be applied here. In the rest of this section I will summarise my own ideas in this respect — these are simple techniques which I believe have a clear justification that coheres well with the entropy-based approach of the last section.⁴² First I shall deal with the case where new causal information comes to be known. After this I shall address the following questions. What sort of information should one collect in order to best refine the network? How one can limit the complexity of the network?

6.1 CAUSAL INFORMATION

Suppose our agent X finds out that C_i causes C_j . I suggest that she should just add an arrow from C_i to C_j to her initial causal graph (if there is no arrow there already), and she should ensure her specifying probabilities $p(c_j|d_j)$ take this new parent into account. There are two possible justifications of this *adding-arrows* strategy. One can apply the arguments of the last section. If X learns of the new causal link and the corresponding probabilities then her background knowledge now includes an extended causal graph and probability specification, in which case she should maximise entropy by adopting the new Bayesian network formed by adding the arrow and the specifiers.

The second possible justification relies on the dependence principle⁴³ as opposed to causal irrelevance, as follows. Suppose we start off with Bayesian network (G, S_G) , where G is X 's causal graph and S_G is her associated probability specification, whose entries we shall assume agree with the objective probabilities $p^*(c_i|d_i)$. Then we add an arrow from C_i to C_j and change the specified probabilities to give a new network (H, S_H) . We measure the improvement of the new network over the old by how much closer its induced probability function p_H is to the objective probability function p^* than p_G , according to the usual measure of distance between probability functions,

⁴¹See [Jordan 1998] and [Buntine 1996] for good surveys.

⁴²Some related work: the Kutató algorithm of [Herskovitz 1991] also has an entropy-based justification. However it involves minimising entropy and poses significant computational problems in the worst case. [Jitnah 1999] employs mutual information as I do, but as a technique for probabilistic inference given a Bayesian network rather than a means for deriving the network itself.

⁴³Recall that the dependence principle says that a direct cause changes the probability of its effect conditional on the effect's other causes.

cross entropy. Then we have the following facts:

IMPROVEMENT OF ADDING ARROWS

- (i) the new network is no worse a network than the initial network;
- (ii) the new network is a better network if and only if C_j is probabilistically dependent on C_i , conditional on C_j 's other parents D .

In particular, if the dependence principle holds then the fact that C_i is a cause of C_j entails that the two nodes are conditionally probabilistically dependent and thus that the probability distribution represented by the new network is closer to the target objective distribution than that of the old network: we are justified in adding an arrow from C_i to C_j .

PROOF: For simplicity (but without loss of generality as we shall see shortly) we shall assume that p_G and p_H are strictly positive over the *atomic states* $c_1 \wedge \dots \wedge c_N$.

For (i) we need to show that $d(p^*, p_H) - d(p^*, p_G) \leq 0$, where d is cross entropy distance. So,

$$\begin{aligned} d(p^*, p_H) - d(p^*, p_G) &= \sum_s p^*(s) \ln \frac{p^*(s)}{p_H(s)} - \sum_s p^*(s) \ln \frac{p^*(s)}{p_G(s)} \\ &= \sum_s p^*(s) \ln \frac{p_G(s)}{p_H(s)}, \end{aligned}$$

where the s are the atomic states, and bearing in mind that $p_H(s) > 0$. Now for real $x > 0$, $\ln(x) \leq x - 1$. By assumption $p_G(s)/p_H(s) > 0$, so

$$\begin{aligned} \sum_s p^*(s) \ln \frac{p_G(s)}{p_H(s)} &\leq \sum_s p^*(s) \left[\frac{p_G(s)}{p_H(s)} - 1 \right] \\ &= \sum_s p^*(s) \frac{p_G(s)}{p_H(s)} - 1, \end{aligned}$$

and thus we need to show that

$$\sum_s p^*(s) \frac{p_G(s)}{p_H(s)} \leq 1.$$

Now since we are dealing with Bayesian networks,

$$\frac{p_G(s)}{p_H(s)} = \frac{\prod p^*(c_k | d_k^G)}{\prod p^*(c_k | d_k^H)},$$

for each literal c_k consistent with s , where d_k^G is the state of the parents of C according to G which is consistent with s , and likewise for d_k^H . But H is

just G but with an arrow from C_i to C_j , so the terms in each product are the same and cancel, except when it comes to literals c_j involving node C_j . Thus

$$\frac{p_G(s)}{p_H(s)} = \frac{p^*(c_j|d_j^G)}{p^*(c_j|d_j^H)} = \frac{p^*(c_j|d)}{p^*(c_j|c_i \wedge d)},$$

where we just let d be d_j^G and c_i the remaining literal in d_j^H . Substituting and simplifying,

$$\begin{aligned} \sum_s p^*(s) \frac{p_G(s)}{p_H(s)} &= \sum p^*(c_i \wedge c_j \wedge d) \frac{p^*(c_j|d)}{p^*(c_j|c_i \wedge d)} \\ &= \sum p^*(c_j|d) p^*(d|c_i) p^*(c_i). \end{aligned}$$

Consider the new set of variables $\{C_i, C_j, D\}$ where C_i and C_j are as before and D takes as values the states of the parents of C_j according to G . Form a Bayesian network T incorporating the graph $C_i \longrightarrow D \longrightarrow C_j$ (with specifying probabilities determined as usual from the probability function p^*). Then since T is a Bayesian network, $\sum p^*(c_j|d) p^*(d|c_i) p^*(c_i) = \sum p_T(c_i \wedge c_j \wedge d) = 1$ by the additivity of probability. Thus $\sum_s p^*(s) p_G(s)/p_H(s) = 1$ so $d(p^*, p_H) - d(p^*, p_G) \leq 0$, as required.

Let us now turn to (ii). From the above reasoning we see that

$$d(p^*, p_H) - d(p^*, p_G) < 0 \Leftrightarrow \ln \frac{p_G(s)}{p_H(s)} < \frac{p_G(s)}{p_H(s)} - 1$$

for some atomic state s . But $\ln x < x - 1 \Leftrightarrow x \neq 1$, and

$$\frac{p_G(s)}{p_H(s)} \neq 1 \Leftrightarrow \frac{p^*(c_j|c_i)}{p^*(c_j|c_i \wedge d)} \neq 1 \Leftrightarrow p^*(c_j|c_i \wedge d) - p^*(c_j|d) \neq 0,$$

where the c_i, c_j, d are consistent with s . Therefore, $d(p^*, p_H) - d(p^*, p_G) < 0$ if and only if there is some c_i, c_j, d for which the conditional dependence holds.

The assumption that p_G and p_H are positive over atomic states is not essential. Suppose p_H is zero over some atomic states. Then in the above,

$$\begin{aligned} \sum_s p^*(s) \ln \frac{p_G(s)}{p_H(s)} &= \\ \sum_{s:p_H(s)>0} p^*(s) \ln \frac{p_G(s)}{p_H(s)} &+ \sum_{s:p_H(s)=0} p^*(s) \ln \frac{p_G(s)}{p_H(s)}. \end{aligned}$$

The first sum on the right hand side is ≤ 0 as above. The second sum is zero because each component is, as we shall see now. Suppose $p_H(s) = 0$.

Then $\prod_{k=1}^N p^*(c_k | d_k^H) = 0$ so $p^*(c_k \wedge d_k^H) = 0$ for at least one such k , in which case $p(s) = 0$ since by the axioms of probability, $p(u) = 0 \Rightarrow p(u \wedge v) = 0$. Now in the sum read $p^*(s) \ln p_G(s)/p_H(s)$ to be $p^*(s) \ln p_G(s) - p^*(s) \ln p_H(s)$. In dealing with cross entropy by convention $0 \ln 0$ is taken to be 0. Therefore $p^*(s) \ln p_G(s)/p_H(s) = 0 \ln p_G(s) - 0 = 0$. The same reasoning applies if p_G is zero over some atomic states. Likewise, if $p^*(s)$ is zero then $p^*(s) \ln p_G(s)/p_H(s)$ is zero too. \square

This justifies the adding-arrows approach if X learns of a new causal link amongst the current variables. If she learns of a new variable C_{N+1} that is causally related to one or more of the other variables, and she also learns the probabilities $p(c_{N+1} | d_{N+1})$, then we can apply the above argument (or equally the arguments of §5) to show that X 's new network should be constructed from her old network by adding the new node and causal arrows to her graph and the new probabilities to her specification.

Finally note that the above argument only requires that the added arrow links conditionally probabilistically dependent nodes. As we have discussed in §2, nodes need not be causally related to be probabilistically dependent. Therefore, if our agent is presented with information to the effect that two nodes are conditionally dependent, she is justified in adding the corresponding arrow to her network, regardless of whether those nodes are causally related. But as a result of this generalisation, the graph in the agent's Bayesian network need no longer be causally interpreted: the Bayesian network becomes an abstract tool for representing a probability function.

6.2 MUTUAL INFORMATION

We now have a strategy for changing the network when causal information or other probabilistic dependencies are presented to the agent. But is there a strategy for *seeking out* a good arrow to add? By adding arrows we increase both the size of the specification required in the Bayesian network (the *space complexity*) and the time taken to calculate probabilities from the network (the *time complexity*) — is there a means of limiting these complexities to prevent the network from becoming impractical? I shall address both these questions in this section.

The key to limiting complexity consists in finding constraints \mathcal{C} such that Bayesian networks satisfying \mathcal{C} have acceptable complexity, and then ensuring that (i) the current network satisfies \mathcal{C} , and (ii) an arrow is only added to the current network if the resulting network continues to satisfy \mathcal{C} . Consider by way of example the following constraints.

- \mathcal{C}_1 : no node has more than K parents, for some constant K . This bound on the number of parents serves to restrict the space complexity of a Bayesian network. For instance if $K = 0$ then the discrete network (no arrows) is the only available network, if $K = 1$ then all networks satisfying \mathcal{C}_1 have graphs that are forests, and if $K = N - 1$ there is no restriction at all on the networks. It is easy to see that if all variables are binary, the complexity of a network satisfying \mathcal{C}_1 is less than or equal to $(N - K + 1)2^K - 1$, a value that is linear in N .
- \mathcal{C}_2 : the Bayesian network has space complexity of at most κ . Now if $\kappa = N$ the only network to satisfy \mathcal{C}_2 is the discrete network and if $\kappa = 2^N - 1$ any network satisfies the constraint. Depending on the problem in hand and available resources we will want to choose an appropriate value for κ or K which balances the range of networks available with their complexity.
- \mathcal{C}_3 : the graph is singly-connected. Having a singly connected graph ensures that the Bayesian network can be used to calculate required probabilities efficiently (in time linear in the number of nodes N). Note however that a singly-connected network can have space complexity up to $2^{N-1} + N - 1$ on binary-valued nodes, so in practice this constraint may best be used with another which limits space complexity.

In sum, if we fix some constraints \mathcal{C} the goal then is to find a *constrained network* (a Bayesian network satisfying \mathcal{C}) which gives a good approximation to the target objective distribution p^* (using cross entropy as a measure of degree of approximation).

We shall associate a *weight* with each arrow in a Bayesian network as follows. In order to weigh the arrows going into a node C_i we enumerate the parents of C_i as D^1, \dots, D^k . Then for $j = 1, \dots, k$ we weigh the arrow from D^j to C_i by the *conditional mutual information*,

$$I(C_i, D^j | \{D^1, \dots, D^{j-1}\}) = \sum_{c_i, d, d^j} p^*(c_i \wedge d \wedge d^j) \log \frac{p^*(c_i \wedge d^j | d)}{p^*(c_i | d)p^*(d^j | d)},$$

where d ranges over the states $d^1 \wedge \dots \wedge d^{j-1}$. Then:

MAX-WEIGHT APPROXIMATION

The network subject to constraints \mathcal{C} which affords the closest approximation to p^* (according to the cross entropy measure of distance) is the network satisfying \mathcal{C} whose arrow weights are maximised.

PROOF: The distance between the probability function p determined by X 's Bayesian network and the target function p^* is

$$d(p^*, p) = \sum_s p^*(s) \log \frac{p^*(s)}{p(s)}$$

$$= \sum_s p^*(s) \log p^*(s) - \sum_s p^*(s) \log \prod_{i=1}^N p^*(c_i | d_i)$$

where the c_i and d_i are consistent with s ,

$$\begin{aligned} &= \sum_s p^*(s) \log p^*(s) - \sum_s p^*(s) \sum_{i=1}^N \log p^*(c_i | d_i) \\ &= \sum_s p^*(s) \log p^*(s) - \sum_s p^*(s) \sum_{i=1}^N \log \frac{p^*(c_i \wedge d_i)}{p^*(c_i) p^*(d_i)} - \sum_s p^*(s) \sum_{i=1}^N \log p^*(c_i) \\ &= -H(p^*) - \sum_{i=1}^N I(C_i, D_i) + \sum_{i=1}^N H(p^*|_{C_i}) \end{aligned}$$

where $H(p^*)$ is the entropy of function p^* , $I(C_i, D_i)$ is the mutual information between C_i and its parents and $H(p^*|_{C_i})$ is the entropy of p^* restricted to node C_i . The entropies are independent of the choice of Bayesian network so the distance between the network and target distributions is minimised just when the total mutual information is maximised.⁴⁴

Note that

$$\begin{aligned} &I(A, B) + I(A, C|B) \\ &= \sum_{a,b,c} p^*(a \wedge b \wedge c) \left[\log \frac{p^*(a \wedge b)}{p^*(a)p^*(b)} + \log \frac{p^*(a \wedge c|b)}{p^*(a|b)p^*(c|b)} \right] \\ &= \sum_{a,b,c} p^*(a \wedge b \wedge c) \log \frac{p^*(a \wedge b)p^*(a \wedge b \wedge c)p^*(b)p^*(b)}{p^*(a)p^*(b)p^*(b)p^*(a \wedge b)p^*(c \wedge b)} \\ &= \sum_{a,b,c} p^*(a \wedge b \wedge c) \log \frac{p^*(a \wedge b \wedge c)}{p^*(a)p^*(c \wedge b)} = I(A, \{B, C\}). \end{aligned}$$

By enumerating the parents D_i of C_i as D^1, \dots, D^k , we can iterate the above relation to get

$$I(C_i, D_i) = I(C_i, D^1) + I(C_i, D^2|D^1) +$$

⁴⁴This much is a straightforward generalisation of the proof of [Chow & Liu 1968] that the best tree-based approximation is the maximum weight spanning tree.

$$I(C_i, D^3 | \{D^1, D^2\}) + \dots + I(C_i, D^k | \{D^1, \dots, D^{k-1}\}).$$

Therefore,

$$\sum_{i=1}^N I(C_i, D_i) = \sum_{i=1}^N \sum_j I(C_i, D^j | \{D^1, \dots, D^{j-1}\}),$$

and the cross entropy distance between the network distribution and the target distribution is minimised just when the sum of the arrow weights is maximised. \square

Note that this result is independent of choice of enumeration of the variables, as can be seen from the proof.

There are various ways one might try to find a constrained network with maximum or close to maximum weight, but perhaps the simplest is a greedy adding-arrows strategy: start off with the discrete graph and at each stage find and weigh the arrows whose addition would ensure that the dag structure and constraints \mathcal{C} remain satisfied, and add one with maximum weight. If more than one best arrow exists we can spawn several new graphs by adding each best arrow to the previous graph, and we can constantly prune the number of graphs by eliminating those which no longer have maximum weight. We stop the algorithm when no more arrows can be added.⁴⁵

Given this algorithm and its justification, we now have answers to our two questions of this section. We seek out a good arrow to add by finding the arrow with maximum conditional mutual information weight. We limit the complexity of the network by imposing constraints on the network.

Thus in stage two of the two-stage methodology we can improve the causal network obtained in stage one by adding arrows — these arrows link causally related variables or more generally probabilistically dependent variables, and a good strategy is to add the weightiest arrow which does not violate constraints on the complexity of the network. The conditional mutual information weighting is a measure of conditional dependence and so in effect the strategy is to add an arrow between two nodes that are most (conditionally) dependent. The resulting graph will not necessarily reflect the true causal relations amongst the variables, and so stage two corresponds more closely to the abstract foundations for Bayesian networks than any causal interpretation.

⁴⁵See [Williamson 2000b] and [Williamson 2000] for analyses of the performance of this algorithm, which turns out to be remarkably effective for a greedy approach.

CONCLUSION

While the independence assumption poses significant problems for a straightforward objective or subjective interpretation of Bayesian networks, independence can be thought of as a means of determining a rational belief function from an agent's background knowledge. Thus Bayesian networks can be given firm foundations by adopting a two-stage approach, whereby one first adopts a subjective causal interpretation which may then be dropped as the network is refined in order to better approximate a target objective probability function. These foundations appeal to information-theoretic notions and assumptions about causality which are somewhat less contentious than the independence assumption. Stage one is justified by maximum entropy considerations while an adding-arrows strategy for stage two can be justified by minimising cross entropy relative to the objective distribution. This approach is not subject to many of the problems that beset the objective or subjective interpretations considered in §2 and §3: we do not need to worry about individuation of variables, and stage two can be used to compensate for the presence of accidental and extra-causal dependencies and any discrepancies between the subjective network and an objective causal network. The advantage over the abstract approach is that we don't require a database of past case data to determine a network — stage one makes use of causal and probabilistic background knowledge. The two-stage methodology can be viewed as a way of integrating background knowledge (including qualitative causal knowledge) with machine learning techniques (of which the adding-arrows strategy is one example).⁴⁶

REFERENCES

- [Arntzenius 1992] Frank Arntzenius: 'The common cause principle', *Philosophy of Science Association* 1992 (2), pages 227-237.
- [Binder et al. 1997] John Binder, Daphne Koller, Stuart Russell & Keiji Kanazawa: 'Adaptive probabilistic networks with hidden variables', *Machine Learning* 29, pages 213-244.

⁴⁶Thanks to David Corfield, Donald Gillies and Jeff Paris for helpful comments, and the UK Arts and Humanities Research Board for funding this research.

- [Buntine 1996] Wray Buntine: ‘A guide to the literature on learning probabilistic networks from data’, *IEEE Transactions on Knowledge and Data Engineering* 8(2), pages 195-210.
- [Butterfield 1992] Jeremy Butterfield: ‘Bell’s theorem: what it takes’, *British Journal for the Philosophy of Science* 43, pages 41-83.
- [Cartwright 1989] Nancy Cartwright: ‘Nature’s capacities and their measurement’, Oxford: Clarendon Press.
- [Chow & Liu 1968] C.K. Chow & C.N. Liu: ‘Approximating discrete probability distributions with dependence trees’, *IEEE Transactions on Information Theory* IT-14, pages 462-467.
- [Fisher 1935] Ronald Fisher: ‘The design of experiments’, Edinburgh: Oliver & Boyd.
- [van Fraassen 1980] Bas C. van Fraassen: ‘The scientific image’, Clarendon Press, Oxford.
- [Grünwald 2000] Peter Grünwald: ‘Maximum entropy and the glasses you are looking through’, *Proceedings of the 16th conference of Uncertainty in Artificial Intelligence*, Stanford University, Morgan Kaufmann, pages 238-246.
- [Hausman 1999] Daniel M. Hausman: ‘The mathematical theory of causation’, review of [McKim & Turner 1997], *British Journal for the Philosophy of Science* 50, pages 151-162.
- [Healey 1991] Richard Healey: ‘Review of Paul Horwich’s “Asymmetries in time”’, *The Philosophical Review* 100, pages 125-130.
- [Herskovitz 1991] Edward Herskovitz: ‘Computer-based probabilistic-network construction’, PhD Thesis, Stanford University.
- [Humphreys 1997] Paul Humphreys: ‘A critical appraisal of causal discovery algorithms’, in [McKim & Turner 1997], pages 249-263.
- [Humphreys & Freedman 1996] Paul Humphreys & David Freedman: ‘The grand leap’, *British Journal for the Philosophy of Science* 47, pages 113-123.
- [Jitnah 1999] Nathalie Jitnah: ‘Using mutual information for approximate evaluation of Bayesian networks’, PhD Thesis, School of Computer Science and Software Engineering, Monash University.

- [Jordan 1998] Michael I. Jordan(ed.): ‘Learning in Graphical Models’, Cambridge, Massachusetts: The M.I.T. Press 1999.
- [Kwoh & Gillies 1996] Chee-Keong Kwoh & Duncan F. Gillies: ‘Using hidden nodes in Bayesian networks’, *Artificial Intelligence* 88, pages 1-38.
- [Lad 1999] Frank Lad: ‘Assessing the foundation for Bayesian networks: a challenge to the principles and the practice’, *Soft Computing* 3(3), pages 174-180.
- [Lemmer 1993] John F. Lemmer: ‘Causal modeling’, in *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence*, San Mateo: Morgan Kaufmann, pages 143-151.
- [Lemmer 1996] John F. Lemmer: ‘The causal Markov condition, fact or artifact?’, *SIGART* 7(3).
- [McKim & Turner 1997] Vaughn R. McKim & Stephen Turner: ‘Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences’, University of Notre Dame Press.
- [Mill 1843] John Stuart Mill: ‘A system of logic, ratiocinative and inductive: being a connected view of the principles of evidence and the methods of scientific investigation’, New York: Harper & Brothers, eighth edition, 1874.
- [Neapolitan 1990] Richard E. Neapolitan: ‘Probabilistic reasoning in expert systems: theory and algorithms’, New York: Wiley.
- [Oliver & Smith 1990] R.M. Oliver & J.Q. Smith: ‘Influence diagrams, belief nets and decision analysis’, Chichester: Wiley.
- [Papineau 1992] David Papineau: ‘Can we reduce causal direction to probabilities?’, *Philosophy of Science Association* 1992 (2), pages 238-252.
- [Paris 1994] Jeff Paris: ‘The uncertain reasoner’s companion’, Cambridge: Cambridge University Press.
- [Paris 1999] Jeff Paris: ‘Common sense and maximum entropy’, *Synthese* 117, pages 73-93.
- [Paris & Vencovská 1997] Jeff Paris & Alena Vencovská: ‘In defense of the maximum entropy inference process’, *International Journal of Automated Reasoning* 17, pages 77-103.

- [Paris & Vencovská 2001] J.B. Paris & A. Vencovská: ‘Common sense and stochastic independence’, this volume.
- [Pearl 1988] Judea Pearl: ‘Probabilistic reasoning in intelligent systems: networks of plausible inference’, Morgan Kaufmann.
- [Pearl 2000] Judea Pearl: ‘Causality: models, reasoning, and inference’, Cambridge University Press.
- [Pearl et al. 1990] Judea Pearl, Dan Geiger & Thomas Verma: ‘The logic of influence diagrams’, in [Oliver & Smith 1990], pages 67-87.
- [Price 1992] Huw Price: ‘The direction of causation: Ramsey’s ultimate contingency’, *Philosophy of Science Association* 1992 (2), pages 253-267.
- [Reichenbach 1956] Hans Reichenbach: ‘The direction of time’, Berkeley & Los Angeles, University of California Press, reprinted 1971.
- [Rolnick 1974] William B. Rolnick: ‘Causality and physical theories’, New York: American Institute of Physics.
- [Salmon 1980] Wesley C. Salmon: ‘Probabilistic causality’, in [Salmon 1998], pages 208-232.
- [Salmon 1984] Wesley C. Salmon: ‘Scientific explanation and the causal structure of the world’, Princeton: Princeton University Press.
- [Salmon 1998] Wesley C. Salmon: ‘Causality and explanation’, Oxford: Oxford University Press.
- [Savitt 1996] Steven F. Savitt: ‘The direction of time’, *British Journal for the Philosophy of Science* 47, pages 347-370.
- [Schlegel 1974] Richard Schlegel: ‘Historic views of causality’, in [Rolnick 1974], pages 3-21.
- [Sober 1988] Elliott Sober: ‘The principle of the common cause’, in James H. Fetzer (ed.): ‘Probability and causality: essays in honour of Wesley C. Salmon’, pages 211-228.
- [Spirtes et al. 1993] Peter Spirtes, Clark Glymour & Richard Scheines: ‘Causation, Prediction, and Search’, *Lecture Notes in Statistics* 81, Springer-Verlag.

- [Spirtes et al. 1997] Peter Spirtes, Clark Glymour & Richard Scheines: ‘Reply to Humphreys and Freedman’s review of ‘Causation, prediction, and search’’, *British Journal for the Philosophy of Science* 48, pages 555-568.
- [Sucar et al. 1993] L.E. Sucar, D.F. Gillies & D.A. Gillies: ‘Objective probabilities in expert systems’, *Artificial Intelligence* 61, pages 187-208.
- [Sundaram 1996] Rangarajan K Sundaram: ‘A first course in optimisation theory’, Cambridge: Cambridge University Press.
- [Williamson 1999] Jon Williamson: ‘Does a cause increase the probability of its effects?’, philosophy.ai report pai_jw_99_d, <http://www.kcl.ac.uk/philosophy.ai>.
- [Williamson 2000] Jon Williamson: ‘A probabilistic approach to diagnosis’, *Proceedings of the Eleventh International Workshop on Principles of Diagnosis (DX-00)*, Morelia, Michoacan, Mexico, June 8-11 2000.
- [Williamson 2000b] Jon Williamson: ‘Approximating discrete probability distributions with Bayesian networks’, in *Proceedings of the International Conference on Artificial Intelligence in Science and Technology*, Hobart Tasmania, 16-20 December 2000, pages 106-114.

INDEX

abstract structures, 3
accidentally correlated, 6
adding-arrows, 36
ancestrally, 31
atomic states, 37

background knowledge, 26
Bayesian network, 2
Bayesian Networks Maximise En-
 tropy, 30

causal extension, 6
causal irrelevance, 28
causal Markov condition, 1, 3
causal restriction, 8, 9
conditional mutual information, 40
constrained network, 40
correlation restriction, 8
cross entropy, 31, 37

dag, 2
dependence, 22
direct method, 3

existence, 5
extra-causal constraint, 6

improvement, 36
Improvement of Adding Arrows, 37
interpreted, 3

knowledge elicitation problem, 4

literal, 2

Max-Weight Approximation, 40
maximum entropy principle, 27

principle of indifference, 27
principle of the common cause, 5
propagation algorithms, 3
 restriction, 8

 screening, 5
 singly-connected, 40
 space complexity, 39
 stage one, 26
 stage two, 26
 state, 2

 time complexity, 39
 two-stage methodology, 26

weight, 40