Appeared in SIGART, vol 7, num 3

The Causal Markov Condition, Fact or Artifact?

John F. Lemmer
Rome Laboratory
525 Brooks Road
Rome, NY 13441
John.Lemmer@rl.af.mil

Abstract

This paper provides *a priori* criteria for determining when a causal model is sufficiently complete to be considered a Bayesian Network, and a new representation for Bayesian Networks shown to be more computationally efficient in a wide range of circumstances than current representations.

Expert Systems for domains in which uncertainty plays a major role are often built from causal models. These models are usually *implemented* using Bayesian Network technology under the often tacit assumption that a causal model *is* a satisfactory Bayesian Network of the domain. If the system produces unsatisfactory results, the causal model is usually deemed inadequate, probably containing insufficient detail.

The assumption that a causal model is an *appropriate* Bayesian Network model is justified by invoking the so called "Causal Markov Condition". In this paper we argue that in many cases it is not the inadequacy of the causal model which produces unsatisfactory results, but rather the inappropriateness of the Causal Markov Condition itself.

In this paper we introduce a new functional model of causality, the Communicating Causal Process model, and analyze the appropriateness of the Causal Markov Condition in light of this model. This analysis yields domain based *a priori* criteria for judging when the Causal Markov Condition does or does not hold, and when a Causal Model is sufficiently detailed that it can be considered a Bayesian Network.

The Communicating Causal Process model also provides the basis for a new representation of Bayes Networks which shown to be more computationally efficient than current representations.

1. BACKGROUND AND INTRODUCTION

Knowledge Engineers building Causal Models implemented as Bayesian Networks face a difficult decision: determining when the model is sufficiently complete to satisfy Bayesian Network assumptions, yet compact enough to be computationally tractable. The major contributions of this paper are to provide domain based tests for sufficient completeness and a more efficient representation.

The next subsection, Background, introduces the concepts underlying graphical causal models and Bayesian Networks. It also shows how the Causal Markov Condition is used to establish an isomorphism between the two. We demonstrate how this isomorphism is often established in ways inconsistent with the modeled domain. This demonstration is likely of interest even to those who are already familiar with Bayesian Networks.

The following subsection discusses why it is important to consider whether the Causal Markov Condition is, as often alleged, a fact of the physical world and our understanding of causality, or is an artifact of the way in which we have constructed our models.

1.1 BACKGROUND

This section introduces graphical causal models, Bayesian Networks, and the Causal Markov Condition in sufficient detail to make the remainder of the paper understandable to those with only a basic background in probability.

This section introduces the concept of an *appropriate* Bayesian Network for a particular causal model, a concept likely to be of interest to all readers. It also introduces graphical causal models, Bayesian Networks, and the Causal Markov Condition in sufficient detail to make the remainder of the paper understandable to those with only a basic background in probability.

The principle ideas to bear in mind throughout the following subsections are:

- A causal model is a model of a *domain*.

- A Bayesian Network is a model of *a probability distribution.*
- The Causal Markov Condition is an assumption often invoked to make a Bayesian Network isomorphic with a causal model.

1.1.1 Graphical Causal Models

A graphical causal model is a graph in which each node represents a domain process of some sort and an edge from a node, *a,* to a node, *b,* represents the relation that *a* is an immediate cause of *b.* The node, *b,* may have more that one immediate cause. Call the set of immediate causes Pa(b).

The notion of "immediate cause" has been a source of difficulty for philosophers throughout the ages. It is sufficient for our purposes here to follow Elby (Elby 92) who interprets Reichenbach to say that for a deterministic causal model, Pa(b) contains the minimal number of processes necessary such that knowledge of the state of each of the process in Pa(b) is sufficient to predict the state of *b.* (Uncertain causal processes will be addressed below.) So if *a* is an element of P(b), it is an immediate cause *b.*

The graph on Figure 1 is an example of a deterministic causal model.

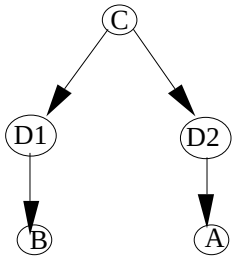


Figure 1: A Causal Model

It is intended to describe certain aspects of a television, and has a long history in the discussion of the Causal Markov condition. C is the on/off switch of the television which can activate both the power bus to the sound, D1, and the bus to the video, D2. The sound, B, reacts to the power on D1 to produce sound; likewise the video reacts to D2.

In a graphical causal model, each 'effect' node is usually augmented with a function in which the effect is the dependent variable, and its immediate causes are the independent variables. For example, in Figure 1, D1 would be augmented with the function D1 = f(C). If the diagram included an edge from D2 to B, then B would be augmented with some function B = g(D1, D2).

The model in this form clearly embodies Reichenbach's claim that immediate causes screen their effects from indirect causes: the state of B can be determined without knowledge of the state of C. But Reichenbach, at least as interpreted by Pearl, has made a stronger statement. Reichenbach has said that immediate causes screen effects from *all* other nodes, not just from indirect causes such as C with respect to B. This stronger claim lies at the heart of the Causal Markov Condition, as we shall soon see. It is this stronger form of this statement with we shall take issue below.

Graphical causal models can be either deterministic on non-deterministic. If the model is deterministic, then the dependent variable, the effect, always takes on the same value for any particular combination of values of its immediate causes. If the model is non-deterministic, the effect can take on with some probability any value in its domain for each combination of its causes[1]. If we represent this non-deterministic function by a table like that in Table 1.

B/D1, D2	00	01	10	11
0	.8	.5	.2	.1
1	.2	.5	.8	.9

Table 1: B, A Non-Deterministic Function

According to this table, if both D1 and D2 have value 1, then there is a .9 probability that B will have value 1, and a .1 probability that it will have value 0.

Non-deterministic causal models augmented with functions like those in Table 1 are syntactically identical with Bayesian Networks. This identity arises because functions such as these are identical with the conditional probability tables of a Bayesian Network. This syntactic identity is further explained in the next subsection.

But is it an appropriate Bayesian Network for the domain? We argue below that it is not. As we shall show below, a network such as this already embodies the Causal Markov Condition and that the Causal Markov Condition places some strong restrictions on intuitive notions of causality.

1.1.2 Bayesian Networks

Bayesian Networks are graphical models of the chain rule representation of a probability distribution. Domain based information, when appropriate, is used to simplify the graphical structure of the model thereby restricting the family of distributions which can be represented.

Any probability distribution over *n* variables can be represented by the chain rule

$$p(x_n, x_{n-1}, \ldots x_1) = p(x_n / x_{n-1}, x_{n-2}, \ldots, x_1)\, p(x_{n-1} / x_{n-2}, x_{n-3}, \ldots, x_1) \ldots p(x_1)$$

which follows directly from the definition of conditional probability

$$p(x/y) = \frac{p(x, y)}{p(y)}$$

A Bayesian Network for this completely general representation of an n-variate probability distribution is a directed graph in which

- there is a one to one correspondence between nodes and variables in the distribution.
- for each factor on the right, an edge appears in the graph from each node appearing in the denominator to the node in the numerator.

[1] This definition of a non-deterministic causal model is different than Pearl's [Pearl, 96]. According to Pearl, non-determinism arises not from the process itself, but rather from exogenous random variables appearing in an otherwise deterministic function.

- each node is augmented with a conditional probability table corresponding to the factor in which the node's corresponding variable appears in the numerator.

The full chain rule representation over all the variables appearing in the television model of Figure 1 is

$$p(B, A, D2, D1, C) = p(B/A, D2, D1, C)\, p(A/D2, D1, C)\, p(D2/D1, C)\, p(D1/C)\, p(C)$$

The graph of the corresponding Bayesian Network is shown in Figure 2.

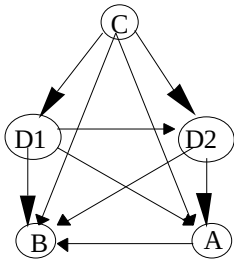


Figure 2: Complete Television Bayes Network

In many domains there are reasons to assume that some "denominator" events are unnecessary. For example in the television model, we may be willing to assume that P(A/D2,D1,C) = p(A/D2), i.e. once we know that power is available from bus D2, knowing the state of the switch provides no further information about the probability a picture appearing. When we make such an assumption it is reflected in the graph by dropping edges corresponding to the 'irrelevant' denominators. In this example, the edges <C, A> and <D1, A> would be dropped.

Let us examine why in the television example we might be willing to assume that P(A/D2,D1,C) = p(A/D2) and drop the corresponding edges from the graph. This examination will ultimately lead to a precise understanding of the Causal Markov assumption itself. We can argue that if we know the state of D2, how the state of D2 was caused provides no further information useful in determining the state of A. Indeed this is implicit in the definition of a direct cause. Likewise we have no need of knowledge of D1 in order to predict A, given knowledge of D2.

We can apply the same arguments to drop edges <C,B>, <D2,B>, and <A,B> from the graph. The new Bayes Network resulting from these domain considerations is shown in Figure 3. But this network is still not isomorphic to our causal model. Can't we apply the same reasoning as above to that pesky edge <D1,D2>?

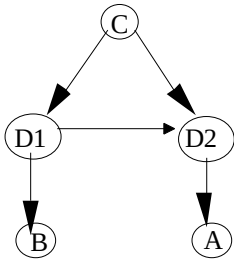


Figure 3: Television Bayes Network Reduced by Domain Considerations

We argue in the next subsection that the same reasoning *cannot* be applied to edges such as <D1,D2>. The difference between this edge and the other edges we did eliminate is that both D1 and D2 share a common independent variable in their causal function. In this case, the event, C, is an immediate cause of both D1 and D2.

1.1.3 The Causal Markov Condition

We feel that the Causal Markov Condition is too general an assumption and that its full generality cannot be defended in *domain* terms. Recall that even such fundamental authorities as Reichenbach were forming conclusions based on domain considerations (Thermodynamics in Reichenbach's case); likewise our discussion below will draw on domain ideas.

Spirtes et al. give the following definition of the Causal Markov condition;

> "Let G be a causal graph with vertex set **V** and *P* be a probability distribution over the vertices in **V** generated by the causal structure represented by *G*. *G* and *P* satisfy the Causal Markov condition if and only if for every W in **V**, W is independent of **V**\(**Descendants**(W).union.**Parents**(W)) given **Parents**(W).

If we assume the Causal Markov condition holds, we have a license for eliminating not only the edges already eliminated by the considerations in Section 2.2.2, but also the edge <D1,D2> in Figure 3:

The justification to remove <D1,D2> is as follows: G is the graph of Figure 1, our causal model of the television. Let W = {D2}; then **V**\(**Descendants**(W).union.**Parents**(W)) = {D1, B} and **Parents**(W) = {C}. Then in particular p(D2/D1,C) is equal by assumption to p(D2/C) so that we are entitled to drop the edge <D1,D2> from the Bayesian Network of Figure 3.

But we claim that assuming p(D2/D1,C) must equal p(D2/C) unduly restricts our intuition of what it means for C to be a cause. In particular it unduly restricts the capabilities of the causal process, C. In particular C cannot be a process that acts in such a way as to induce correlations in its effects. If the uncertainty which is being modeled is conceived of as arising solely from effects sometimes failing to obey "causal commands" issued to them by the cause, then C not inducing correlations among its effects makes sense. But if some of the uncertainly arises from the cause probabilistically, rather deterministically, issuing commands, there is no reason to assume causes cannot induce correlations. We further examine this claim shortly. But first, we note the position of this claim relative to Reichenbach's so called 'dictum'

The position that a single causal process can induce correlations among its effects does not violate Reichenbach's dictum that there is no correlation without causation. The Causal Markov Condition is simply a stronger statement than Reichenbach's. The correlations we discuss above *are* induced by a cause. What this position does violate is the notion, inherent in the Causal Markov Condition, that correlations remaining after conditioning on all known causes must be induced by other unknown *latent* causes. We will show below that the introduction of such latent causes is often no more than a mathematical device to preserve the Causal Markov Condition, and that causes so introduced can be ontologically meaningless in the modeled domain.

Do single causes exist which can induce correlations among their effects? We will provide a notional example and then argue that this example is a good analogy to certain types of 'human' causes.

Suppose that the switch, C, in the television example is implemented in the following way: When C is turned on, a computer embedded in C samples a two variate distribution over D1 and D2 to determine which of these busses power will be supplied power. If D1 and D2 are correlated in this distribution, and there is no reason to suppose that they will not be, then the effects D1 and D2 will be correlated. Of course, if the status of D1 and D2 are incorporated into the state of C, they will be no longer appear in the model as correlated effects violating the Causal Markov Condition. However the remainder of this paper is devoted almost entirely to showing why such incorporation, while often expedient, is neither pragmatically nor ontologically a good idea.

Causes as complex as the switch just described are likely to occur in many situations in which human decision making is involved in the causal process. Thus such complex causes are likely to be important when, for example, actions available to an automatic planner are viewed as causes and plan recognition is an important consideration. Consider the following simplified example:

An army division forming part of a 'front' can be thought of as preparing for one of four activities 1) main attack, 2) holding attack, 3) withdrawal, 4) no movement. Part of the preparation activity involves positioning of air defense assets and of supplies. Such positioning results in effects (probabilistically) observable by the opposing forces. In the case of a main attack moving both forward is most beneficial, and the converse for withdrawal.

The commander of a division has at least two possibly conflicting goals: positioning to best support his impending activities and positioning to deceive his opponent. Thus even if he is about to engage in a main attack, he may not actually choose to move air defenses and supplies forward. He may reason however that if he moves only one forward, the opposition will be alerted anyway, so he may as well move both forward. To his opponents, this would appear that his decision would produce correlated effects.

The division in this case is like a switch with four states. The state of this switch is determined by the division commander's commander. The division commander is the computer which (as viewed by the opponent) samples from a distribution to determine effects.

We feel that whenever Bayesian Network technology is applied to human scale activities, violations of the Causal Markov Condition will be both frequent and important. In the remainder of the paper we discuss both why such violations have so far gone unrecognized, and also how to handle such violations.

1.2 INTRODUCTION

When Bayesian Network technology is used to represent causal models, edges of the network have two different semantics. When the network is constructed, an edge from a to x is interpreted as meaning that a causes x. When probabilistic computations are performed over the constructed model, this same edge is interpreted as requiring certain conditional independencies among events in the model. The justification for assuming that these conditional independencies can be derived from causal relations is termed the Causal Markov assumption.

The Causal Markov assumption is so widely believed and deeply ingrained that, when observed data do *not* exhibit the conditional independencies predicted by a model, it is customary to assume that the domain must contain operative causes which are not yet contained in the model. These additional causes are often termed hidden or *latent* causes. In practice, the relation of these latent causes to the underlying domain is often unclear, but their existence is deemed necessary so that the Causal Markov condition will not be violated. Indeed it will become clear below that these latent causes may often have no relation to the modeled domain and can be better thought of, not as causes, but as mathematical devices for preserving the 'truth' of the Causal Markov assumption.

Thus a Knowledge Engineer building a Bayesian Causal model for a domain in which large amounts of statistical data are not available, must decide when his model is complete enough that the Causal Markov condition will be satisfied. A database 'miner' must decide when 'discovered' causes are significant to the domain. This paper provides criteria on which to base this decision.

The thesis of this paper is that the Causal Markov Assumption, rather than being a fact which is somehow associated with the nature of causality itself, may instead be an artifact of the methods by which causal models have been built and of the way in which data have been analyzed.

The Causal Markov Assumption has been used with such success in model construction and is so unrefuted by data that Spirtes, Glymour, and Scheines [Spirtes 93] have stated:

> "Any persuasive case against The (Causal Markov) Condition would have to exhibit macroscopic systems for which it fails and find some powerful reason why we should think the macroscopic natural and social systems for which we wish causal explanations also fail to satisfy the condition. It seems to us that no such case has been made."

This paper will make just such a case by providing the required "powerful reason." It will not "exhibit the macroscopic systems for which the condition fails," but will suggest approaches by which macroscopic systems may be (re)analyzed so as to detect failures of the Condition. In the conclusion, we will summarize the facts presented below to substantiate our claim of having provided a "powerful reason." We hypothesize, subject to the results of future data analysis, that failures of the condition are not rare. We will suggest how these failures may have gone unrecognized, masked as they may have been by the modeling and analysis techniques which have unwittingly begged the question of the truth of the Condition.

The results of this paper in no way negate work already done based on the Condition. Even if the reader chooses to reject our claim to have found a powerful reason for doubting the applicability of the Condition, the discussion below can be viewed as a continuing refinement of Causal Modeling, offering the potential for both deeper and more intuitive insights into the modeled domain, while also providing improved computational efficiency. It is hoped deeper and perhaps more intuitive insights

will result from more precise representation of modeled domains. Improved efficiency results from elimination of significant redundancy in current techniques used to preserve the Condition in situations where it initially appears to fail.

The powerful reason for doubting the applicability of the Condition to real macroscopic systems will be derived from a simple, plausible, though as yet empirically untested, physical model of the mechanics of "causal influence." This model proposes that causes and effects are both physical processes which communicate with each other by signals generated within the cause and received within the effect. The work here does not *prove* that signals exist. It hypothesizes that signals may exist, explains the benefits of recognizing them, and suggest how they might be found in real domains. Thus this work is theoretical in the same sense as are quantum theories predicting the existence of sub-atomic particles and showing how to detect them.

We call the model from which the powerful reason for doubt is derived the "Communicating Causes" model. This model is described in Section 2. It is in this section that the model is introduced in domain terms followed by discussion of a number of well known examples of counter-examples to the Condition. Following this a formal model is defined for representing Communicating Causal Processes.

In Section 3, we show why, under the assumptions of the Communicating Causes model, effects may be expected often to be correlated. This section begins with a discussion of the nature of signals, and then shows how signals can induce correlation among effects.

Section 4 illustrates some implications of the Communicating Causes model. In this section we show that modeling correlations by introducing extra (non-causal) edges is inappropriate under the assumption of the Communicating Causes Model. Extra edges are inappropriate because the model parameters necessary to capture the correlations are not stationary with respect to priors of the root events.

In Section 5 we show that the Communicating Causes model can offer significant computational efficiencies. One view of the Communicating Causes Model is that it is simply an assessment tool for building causal models and checking the Condition. But in Section 5 we show that when the Condition fails because of interactions which can be modeled as communicating causes, the model also offers computational efficiencies. These efficiencies are apparent relative to the usual method for handling failures of the Condition, namely the introduction of latent causes. It is important to note that we do not claim that every apparent need for a latent cause is better modeled using communicating causes. But where an apparent latent cause is better modeled in this way, large computational savings can result.

2. COMMUNICATING CAUSAL PROCESSES

In the standard Bayesian Network representation of causal relations, a causal process is modeled as a process which observes the state of each of a number of other processes (its causes), and stochastically enters some state of its own. We will extend this model by providing more detail about the nature of the state of a process, and how this state is observed by other processes.

2.1 INFORMAL MODEL

The basic difference between the standard model and the model we are proposing is that in our model an effect event need not directly observe the state of its causes. Instead it observes some 'output' generated by each of its causes. Moreover, a cause may generate a variety of different outputs, only some of which are observable by each of its various effects. These outputs we term signals.

The notion behind our model is that a cause-effect relation in a model corresponds to a communication channel in the domain. In any real world cause-effect relation, some sort of signal is probably transmitted between the causing process and the effected process. It is with the concrete notion of signals that we provide a physical interpretation for the abstract and undefined term "causal influence" often used in discussing causal models. The initiation and broadcast of a signal is under the control of the causing process and is an output of this process. The reception and response to a signal is under control of the effect process. Clearly an effect must be able to receive and respond to at least one of the signal types generated by each of its causes but it need not receive and respond to more than one.

The structure of a standard Bayesian Network causal model suggests that an effect has perfect information about the state of each of its associated causes and that all uncertainty is associated with how the effect will respond to this (perfect) state information. However, it is unlikely that in real physical processes, except perhaps in the most simple cases, effects ever directly and perfectly observe the state of their causes. Instead, it is much more likely that the cause generates some signal which can be sensed by the effect. Thus the physical realization of a cause-effect relation is likely to be some form of communication channel. The cause (probabilistically) generates and transmits some type of signal. The effect senses this signal and responds to it in some way.

Because of our view that the physical realization of a cause-effect relation is a communication channel, our model of causal effect relations explicitly represents two types of uncertainty: first it represents the uncertainty that a cause will actually generate a particular signal given that it is in a particular state, and, second, the uncertainty that the effect will sense and respond to this signal. In contrast the standard Bayesian Network when used to model causality seems to model only the uncertainty that an effect will respond to a complete observation of the states of its causes. In current Bayesian Network practice, both the generation and response uncertainties are usually lumped together into a single uncertainty estimate. Moreover, no notion of partial observation of a cause's state exists.

Even if one is willing to grant that the physical realization of a cause-effect relation ought to be modeled as a communication channel, why shouldn't the signal itself simply be modeled as an effect in the traditional Bayesian Network Markovian manner? The crucial reason is that generation of signals is *entirely* under the control of the causing process and that (because of this) there is no *a priori* reason to rule out correlations among different signals generated by the same process. Note that in our model such correlations can only be ruled out, process by

process, through domain knowledge about how each particular process operates in the physical world. We argue in Section 5 that such correlations may be very common, but have escaped detection because these correlations can also be modeled by the introduction of latent causes.

As an example, consider the cause genotype in the following example adapted from [Heckerman 96]. This model shows genotype (g) influencing both a patient's attitude toward taking (t) a recommendation (r) of a physician and also influencing whether or not a cure (c) occurs.

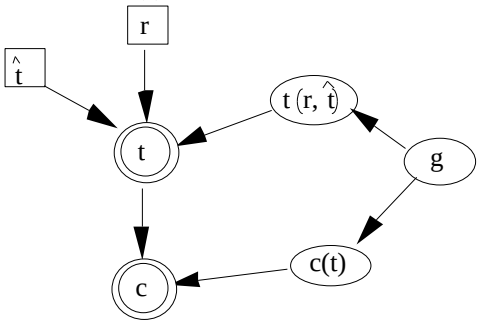


Figure 4: Genotype Influences Cure

It may well be that one part of a gene influences whether or not a patient is a 'complier' or a 'defier', and another part influences how likely it is that a cure will occur. It may well also be that these gene parts often occur together (are correlated)[2]. These gene parts, in our terminology, are signals generated by a causal process. The developmental processes are the receivers of the signal which stochastically respond to these signals and others. The genotype sends correlated signals, and the development processes receive them. Of course it may be that further knowledge of genetics would tell us that the signals are in fact uncorrelated. But the important point is that the correlation (or lack thereof) is a property of the *particular* cause under consideration and not a property of causality in general.

We will show in Section 2.2 below how causes such as genotype can generate correlated effects.

We believe that in many causal models which have been built to date, these signals, *outputs* of the modeled causes, have been treated as though they are just another effect. We suspect that often when the 'fuzzy or' assumption has been used to generate the stochastic response table for an (effect) event, that this event probably ought to be treated as a signal: the essence of the "fuzzy or" assumption is that each cause is entirely capable, through its own agency, of causing the effect. In our terminology such causes can be interpreted as transmitting signals.

2.2 DISCUSSION OF WELL KNOWN EXAMPLES

We believe that our Communicating Causes model of causality makes a convincing case against universal application of the Causal Markov condition. To see just how this case might be made, we will revisit a series of examples covered in [Spirtes 92]. Spirtes *et al*. discuss a number of examples which have been proposed by others as counter examples to the Causal Markov assumption. They reinterpret each of these examples in such a way that the Condition is shown to hold. Here we will reinterpret their reinterpretations in light of our Communicating Causes model.

2.2.1 The Television Set

An example concerning a television set attributed to Davis is quoted in [Spirtes 92]

> Imagine a television set with a balky switch: it usually turns the set on but not always. When the set is on, it produces both sound and picture. Then the probability of a picture given that the switch is on and given sound (is heard) is greater than the probability of a picture given just that the switch is on.

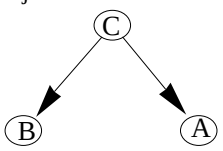


Figure 5: The Balky Television Set

This is shown in Figure 5 where C is the event that the switch is on, B that the sound is on and A that the screen is on. Thus Davis claims that $p(B / C) < p(B / A \wedge C)$. Spirtes *et al*. claim this apparent correlation is due to an unmodeled cause, D, that the circuit is closed and that the correct model is as shown in Figure 6.

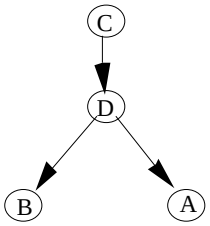


Figure 6: Inside the Balky Television

Their claim is that once we know whether or not the circuit is closed, the sound and picture become probabilistically independent.

Our interpretation of this example is that D is actually the signal transmitted by the causal process embodied in the switch. In our interpretation, sound and picture become independent because they both *receive the same signal*. If they did not receive the same signal, they might still be correlated. Suppose that the sound and picture are each on a separate circuit, each of which is closed by the same switch. Then the correct model becomes as shown in Figure 7 below.

[2] As do the genes for red hair and green eyes

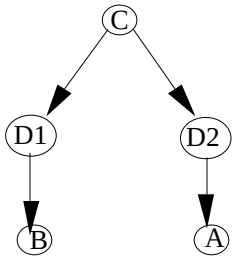


Figure 7: Deeper Inside the Balky Television

It is very possible that C acts on the circuits D1 and D2 in such a way that $p(D1/C) < p(D1/D2,C)$ so that A and B again become correlated.

From a *mathematical* point of view, the situation described in Figure 7 can be handled by at least two methods: either an edge can be introduced between D1 and D2 or else D1 and D2 can be collapsed into a single event, $\hat{D}$, with four states[3]. Both these methods do violence to the notion of a *causal* model. An edge connecting D1 and D2 requires significant mental gymnastics to be interpreted as cause; collapsing D1 and D2 adds structure not present in the real world. Moreover, as we will show below, a single edge will, in general, not be sufficient to model the correlations present in the example. Collapsing D1 and D2 provides information on effects not actually available in the domain. For example consider the probability table which would be associated with B. This table would require four entries for each possible state of B. Each of these entries assumes knowledge of the state of both D1 and D2, but such knowledge is not actually available to B in the physical world. We will show below that, in models with greater complexity, this objection is more than philosophical.

We suggest that the Communicating Causes model offers a criterion to decide how deeply causes must be analyzed. We recognize that it may well be possible to look deeper into the model of Figure 7, identify more (sub)causes and again preserve the Condition. This however will add even further complexity to the model. It may be that even these new causes will need to be analyzed more deeply. This criterion that we offer is that detailed analysis can stop once signals under the control of a single cause have been identified. Then our representation as proposed below is adequate to summarize all relevant information at this level.

2.2.2 The Pool Trick

Spirtes *et al*. discuss the pool trick diagrammed in Figure 8. They define the events A, B, and C such that C means there is a collision of any sort between cue ball and balls one and two, event A means the ball one goes in the pocket L, and event B means that ball two goes in pocket R.

[3] $\hat{D}$ = {⟨~D1, ~D2⟩ ⟨D1, ~D2⟩ ⟨~D1, D2⟩ ⟨D1, D2⟩}

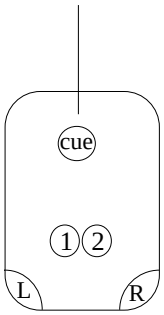


Figure 8: Pool Trick

Clearly if both events C and A occur, we have much more information about the probability of occurrence of event B than if we only know event C occurred. Thus events B and C are correlated even when conditioned on the cause C.

Spirtes *et al*. claim that this is because the state of C does not contain enough detail; perhaps momentum should have been included in the state of C for example. They say "The example simply reflects a familiar problem in real data analysis that arises whenever … distinct values of a variable are collapsed."

We will show below that when correlations result from signals, that it is always possible to invent a new cause (a latent cause) and expand its state space with sufficient information to preserve the Condition. Thus it is not surprising that more state information for the cue ball would provide conditional independence. The question is whether or not this new variable has a clear relation, (or any relation at all) to the domain being modeled[4]. It does have a defined relation in this example, but there is no reason to assume that the general latent variable, invented to preserve the Condition, will have such a relation. Moreover, we will also show below, that even when it does have such a relation, it is usually far from the computationally best representation.

2.3 FORMAL MODEL

We now define our formal model of a causal process, a model intended to reflect the real world. It is a model of the real world in the same sense theories of Physics are models of the real world. The assumptions justifying our model are based on assumptions about the world rather than on assumptions about mathematics.

A causal process is a triple <I, **c**, O> where I is a set of inputs, **c** is the state variable for the process, and O a set of outputs[5] (An event in a standard Bayes Net is shown below to be a special case of a causal process). Associated with each possible

[4] The ancient Irish once believed that noises in the woods at night were caused by leprechauns. While this explanation may have provided them the conditional independence which they sought, the explanation had little relevance to the real world: Leprechauns could neither be measured nor manipulated!

[5] More formally, I is a set of input *random variables*, $\{I_k\}$, each defined over some domain, d_{i_k}; *s* is a single random variable defined over the range, d_s, and O is a set of output random variables $\{O_k\}$.

combination of inputs is a probability distribution over each of the possible states. This is the probability information normally associated with an event in a Bayesian Network and is the probability that the process will enter a particular state when a particular input combination is present. In addition to the data in a normal Bayes Net, there is also associated with each state of a causal process a probability distribution over each possible *combination* of outputs. This distribution gives the probability of each particular combination of outputs *being generated* by this process when it is in that particular state. There is no similar information in Bayesian Networks. In standard Bayesian Network causal models, causes generate no output signals. Rather, their full state information is assumed available to each potential effect.

The elements of both the sets I and O have no analog in the traditional Bayes Net formulation. We term the elements of these set 'signals.' They are intended as an abstraction of how a physical cause triggers physical effects. In the balky television model, the signal, symbolized as D in Figure 4, is appearance of voltage on the bus. The causal process, the switch of the balky television, completely determines whether or not this signal can be generated. The sound generating process determines how it responds to the signal. Signals remain unspecified in traditional Bayes Nets where a causal link simply means that an effect somehow observes the state of the cause.

At this point of our discussion, it may not be yet clear to the reader why the signal either cannot or should not be modeled simply as another event in a Bayes Net. Why signals should be explicitly modeled is the heart of this paper and hopefully the reason will become clear in what follows.

If our model is restricted so that the output set, O, contains just one element, and if the value of this element is always in one to one correspondence with the state, and if the output is always generated whenever a state is entered, then our model becomes equivalent to the standard Bayesian Network formulation. The essence of the standard formulation is that each process can perfectly observe the state of each of its causes. With this restriction on the output set, the state is perfectly observable in our model also. But we believe that in many important real situations, effect events rarely perfectly observe the state of their causes. Moreover in many such situations different effects often observe different information about the same cause.

It is important that a Knowledge Engineer choosing to use the standard Bayes Network formulation realizes (s)he is assuming that perfect and identical state information is available to each and every effect connected to any particular cause. Note that *the validity of this assumption is a property of the domain, not of the mathematics used to model the domain.* Nor is the validity of this assumption a property of any general notion of causality, the so-called "Reichenbach Dictum" not withstanding! Thus validity must be established within the domain context every time a new domain is modeled.

3 HOW EFFECTS BECOME CORRELATED:

In this section, we will show by example how causes, through the agency of signals, can induce correlations among their effects. We will define signals and show how they potentially produce correlated effects. We show in Section 4 why adding "extra edges" to account for these correlations is not appropriate. We show in Section 5 why adding 'latent' causes to account for them is both inappropriate and inefficient.

3.1 Signals

The essence of a signal which makes it an event deserving special representation derives from its (domain dependent) relationship with its cause. A signal is a physical event of some variety which can be *totally* generated by a single cause. A cause may generate more than one signal type and these types are potentially correlated. Correlations cannot be used because the cause is in *complete* control of the signals which it generates. Only domain considerations can rule out correlations. More than one type of cause can generate a particular signal but only one is *required.* Multiple different causes of the same signal are possible and discussed in detail in Section 5.

It is the major conjecture of this paper that signals do exist in nature. However, as will be explained in Section 5, their presence has been masked by the device of latent causes. We will also see in Section 5, that explicit modeling of signals offers many potential advantages.

Mathematically, a signal can be viewed as a communication channel for *partial* information about the state of the cause. Mathematically, signals are related to Hidden Markov Models. As an example, suppose the balky television switch does control two separate circuits. The complete state of the switch includes a description of the closure condition of both circuits. However, physically, neither the sound nor the picture has access to full state information; rather each has access only to information about the voltage on its own wire.

Because a cause controls the signals which it generates, there is no reason to suppose that if it generates more than one, it generates them so that they are probabilistically independent. In Section 5 we will argue that latent variables in many cases may be no more than a mathematical device for modeling causes generating multiple signals, modeling them in a way which preserves the Causal Markov Condition. We will show that in such cases use of this mathematical device adds significant unneeded complexity to causal models.

Spirtes *et al*. might argue that the whole issue is moot, because correlated effects have virtually never been observed which cannot be explained through latent causes. But we suspect, and will argue, again in Section 5, that many "discovered" latent cause are better represented by the signal processing model. The representation through signals is better both in that it is a more faithful representation of the underlying reality, and that it can avoid significant amounts of modeling complexity.

At this point, it is worth pointing out to the concerned reader that abandoning the Condition and adopting the Communicating Causes model will *not* lead to non-computability. Sampling and simulations techniques can be easily adapted, without increase in computational complexity, that have no reliance on this assumption.

3.2 How Signals Induce Correlations

In this subsection we show how causes generating multiple signals can produce correlated effects. In the next subsection we will show that when correlated effects are produced through the agency of signals, it is not a good idea to model these correlations by introducing (even a doubled ended) edge between these effects.

Consider a particular model such as that shown in Figure 9(a). In this model a process, *a*, can cause two effects, *x* and *y*. In Figure 9(b), the signals, σ, transmitted by the cause and received by the effect, are shown explicitly. In terms of our formal model for a causal processes, we have I = {}, **c** = a, and $O = \{\sigma 3, \sigma 7\}$ for the event, *a* shown in Figure 9.

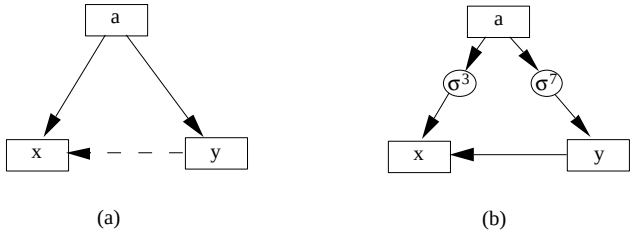


Figure 9: A Simple Causal Model

In the standard Bayesian Network formulation of a causal model, the dotted arrow from *y* to *x* is normally omitted. This is because the nature of causality is assumed to somehow ensure that $p(x\,y / a) = p(x / a)\, p(y / a)$.

In our model we decompose the usual conditional probabilities, such as p(x/a), into two components, the probability of the cause causing the signal and the probability of the effect responding to the signal. The decomposition for p(x/a), given the model in Figure 9(b), is illustrated in equation (1)

$$p(x / a) = p(x / \sigma 3)\, p(\sigma 3 / a) \qquad (1)$$

where $p(\sigma 3 / a)$ is the probability that process *a* will generate the signal and $p(x / \sigma 3)$ is the probability that effect *x* will receive this signal and respond to it. We make a similar assumption with respect to variable *y* also shown in the model in Figure 9(b).

Given our model's assumptions and definitions, we can determine the conditions under which $p(x\,y / a) = p(x / a)\, p(y / a)$ will be true in Figure 9(b). In other words, we can determine the conditions under which the edge from *y* to *x* can be omitted from the figure. By definition:

$$p(xy / a) \stackrel{D}{=} p(xya) / p(a) \qquad (2)$$

so that from the model structure shown in Figure 9(b) and with decompositions of the form of (1)

$$\begin{aligned} p(xy / a) &= \frac{p(x / \sigma 3)\, p(y / \sigma 7)\, p(\sigma 3 \sigma 7 / a)\, p(a)}{p(a)} \\ &= p(x / \sigma 3)\, p(y / \sigma 7)\, p(\sigma 3 \sigma 7 / a) \end{aligned} \qquad (3)$$

If and only if the causal process generates its signals in a probabilistically independent manner, i.e. iff

$$p(\sigma 3 \sigma 7 / a) = p(\sigma 3 / a)\, p(\sigma 7 / a)$$

does (3) become

$$\begin{aligned} p(xy / a) &= p(x / \sigma 3)\, p(\sigma 3 / a)\, p(y / \sigma 7)\, p(\sigma 7 / a) \\ &= p(x/a)\, p(y/a) \end{aligned}$$

where the second equality follows from the decomposition (1). Note that whether or not the causal process generates its signals independently is determined by the nature of the particular process being modeled rather than being determined by the 'nature' of causality.

If and only if the cause generates the signals in probabilistically independent manner are the effects probabilistically independent when conditioned on the cause. Thus the cause's signal generation mechanism determines whether or not its effects are conditionally independent.

3.3 The Test For Model Completeness:

The test for model completeness is based on the assumption that correlations among effects result from the generation of correlated signals within particular causal processes. Thus once each cause in a model has been decomposed to a level where its signals have been identified, there is no need for further decomposition. Recall the definition of signals at the beginning of the section: signals are domain events whose occurrence can be completely determined by its associated cause[6].

Once signals have been identified the model builder can choose the detailed representation. The choices included the traditional form of Bayesian Networks [Pearl 88], or the Communicating Causes model defined above and discussed in more detail in [Lemmer 93]. It is shown below that domain knowledge implies that a causal process is likely to induce correlations among its signals, that the Communication Causes model offers savings in both model building effort and in model representation.

When correlations are expected and traditional Bayes Networks will be used, there are three common ways in which the these correlations can be included in the model. These methods are 1) introduction of (non-causal) edges between events which are thought to be correlated, 2) introduction of an additional (latent) cause which also causes the correlated events, and (3) introduction of an additional cause between the correlation producing cause and its effects.

We show below that methods (1) and (2) are, in general, inappropriate because, under the assumption of signal induced causes, the parameters associated with such added edges and causes are *not modular*[7] but instead vary with the prior (unconditioned) probabilities of other causes in the model.

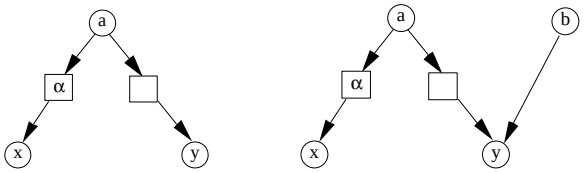
[6] How effects respond to multiple occurrence of the same signal generated by different causal processes or different instances of the same type of causal process is not an issue here.

[7] In Pearl's terms [Pearl 1988], the parameters are not *local*

4 CORRELATED SIGNALS AND ADDITIONAL (NON-CAUSAL) EDGES

As shown above, signals may result in correlations appearing between events such as x and y shown in Figure 9 (a). One technique for dealing with such correlations is to introduce an edge between x and y. But under the assumption of the signal processing model of causality, this will not be sufficient. Edges other than the one from y to x will have to be introduced to preserve model semantics and will further dilute the causal interpretation of the model structure.

Consider the causal model in Figure 10 (a) and the associated probability data in Table 2:



(a) (b)

Figure 10: Causal Models

Note that (b) and (c) in Table 2 are conditional probability tables of the type normally associated with Bayesian Networks, but that (a) is unique to the signaling model of causality because it specifies signals which are correlated due to the operation of their causal process.

Table 2: Signal Type Probability Tables

α,	00	10	10	11
a	.8	.01	.01	.18
ā	1	0	0	0

(a)

	x̄	x
ᾱ	1	0
α	0	1

(b)

	ȳ	y
	1	0
	0	1

(c)

If we assume a prior probability of p(a) = .5 in (a) and use the data in Table 2, we can readily compute p(axy) = .18 , p(ay) = .19 and therefore that p(x/ay) = .948. If to handle this correlation, we insert an arrow from y to x, .948 will be one of the entries in x's conditional probability table.

Suppose now that we introduce another cause, *b*, as shown in Figure 10 (b) and replace Table 2 (c) with the probabilities shown in Table 3.

Table 3: Cause '*b*' Also Interacting with '*y*'

, b	ȳ	y
00	1	0
10	0	1
01	0	1
11	0	1

If we take p(b) = . 5 also, p(ay) takes on a value of .2975 whereas this value was .19 before the addition of b. Thus the value of p(axy) as computed from p(x/ay) p(ay) will increase. But this makes no sense in terms of the underlying causal interpretation of the graph: x is caused only by a. Therefore adding an event b cannot change the probability of events axy jointly occurring. Thus under the assumptions of the signal processing model of causality, what has actually happened is that p(x/ay) has changed: a parameter in a conditional probability table has changed as a result of a change in priors. This parameter is never directly represented in our proposed model so the difficulty this causes in the canonical model is avoided.

In our simple example in Figure 10 (b), the problem can be solved by adding an edge from *b* to *y*. But in doing so, we are certainly moving further and further away from a *causal* model. The original edge between *x* and *y* in Figure 9 certainly did not represent a cause. The second edge seems to be even less a cause, and in fact seems best interpreted as an "anti-cause." Changing the direction of the first extra arrow is no solution either since one need only imagine another event which causes *x* in the same way *b* causes *y* .

It may seem that a solution would lie in explicitly modeling the signals, and having the extra 'correlation' edges between the signals. But the same difficulty recurs: there can be multiple common causes for each signal. In this case, it is probably possible to model each signal separately for each common cause, but then the notion of a *causal* model has definitely been left far behind.

It seems that correlation introduced through the agency of a signal generating causal process simply cannot be adequately modeled by the addition of more edges[8]. However, one can turn to the approach of 'identifying' latent causes. But we will now show that, in many respects, this approach is no more adequate than adding edges.

5 CORRELATED SIGNALS AND LATENT CAUSES

In this section we will demonstrate the representational and computational benefits which accrue when latent causes can be interpreted as correlated signals. We will show that correlated signals can always be modeled as latent causes (though probably not *vice versa)*, but at the expense of greatly increased model complexity. Therefore, whenever latent events can be interpreted as signals, great computation savings should be possible. Whether latent causes ought to be interpreted as signals is, however, a question which can only be answered on a domain by domain basis. We will show that whenever correlated signals are modeled directly as in [Lemmer 93] rather than modeled as latent variables we can achieve much greater representational and computational efficiency[9].

5.1 MODELING CORRELATED SIGNALS AS LATENT CAUSES

In this section we will show how to translate back and forth between the communicating causal processes as defined in the

[8] Double arrowed or not!

[9] In [Lemmer 93] no feasible algorithm was presented for solving models explicitly representing signals. We show below that work by others can be readily adapted to the models explicitly representing correlated signals.

Section 3 and a Bayesian Network in which the Causal Markov assumption holds. It is not surprising that a translation *from* communicating causal processes is possible since Bayesian Networks are sufficiently general to represent any probability distribution. However the translation raises the possibility that latent events may be more a mathematical device than they are a domain modeling tool. Nonetheless this translation does allow traditional d-separation algorithms to be used for prediction and inference, though at the cost of significantly increased model complexity. Another algorithmic approach for prediction and inference is directly applying sampling and simulation techniques to the communicating causes representation.

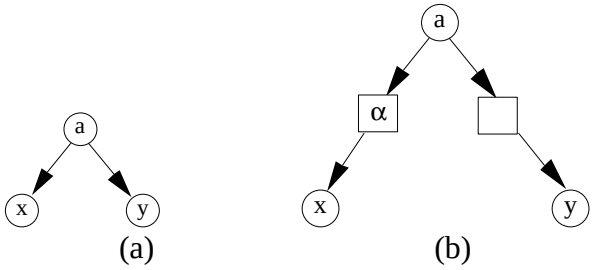


Figure 11: Signal Induced Correlation

To begin understanding the relation between signals and latent causes, consider the example shown in Figure 11(a) where *a* causes both *x* and *y* . Suppose that *x* and *y* are correlated even when conditioned on *a* because *a* generates the signals α and in such a way that they are correlated. These signals are explicitly shown in Figure 11(b) and they induce correlations in *x* and *y* through the mechanism described by equation (3). These are both causal models, but the Causal Markov assumptions cannot be applied. It cannot be claimed that this results from missing information as in the pocket billiards example presented in [Spirtes '93]. From a domain point of view, all relevant information is modeled. But the mathematical device of latent causes, with no necessary relation to the application domain, is capable of rendering the Causal Markov assumption once again applicable!

The mathematical device is to combine the signals generated by a cause into a single latent variable, *L* as shown in Figure 12(a). The states of L are the cross product of the states of the signals modeled by the latent variable as shown in Figure 12(b). With this construction, the cause, *a*, combined with the latent cause, *L*, can be viewed as isomorphic with a causal process, <I, **c**, O>, as defined in Section 2.3. In this particular case, I = {}, $\mathbf{c} \equiv a$ and the states of *c* correspond to all combinations of the signals α and .

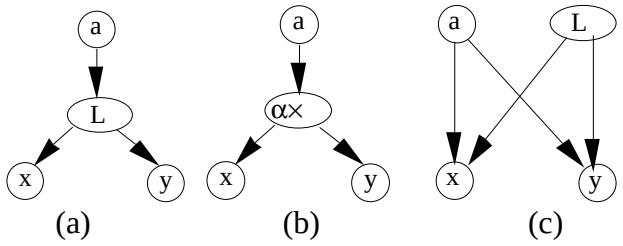


Figure 12: Signals Replaced by a Latent Variable

A common alternative is to introduce the latent variable as shown in Figure 12(c). But note that the table sizes associated with (b) are always less than those associated with (c). Adopting the shorthand that 'a' is the cardinality of the set of discrete states which can be assumed by *a* , etc., the number of table entries required by (b) is α (a + x + y) while the number of entries for (c) is α a(x + y).

There is a second, perhaps even more telling reason for not adopting the latent variable representation shown in Figure 12(c) for modeling causal process induced signal correlations. In the representation of Figure 12(c), the signal correlations produced by *a* are a function not only of the state *a* , but also of the prior probabilities of *a* and *L*. Thus this representation of process induced correlations suffers the same defect as previously noted for representing such correlations by introducing edges between pairs of effects.

The dependence of the correlations on the priors of *a* and *L* is shown by

$$p(xy/a) = \sum_{\alpha} p(x/\alpha\ a)\ p(\alpha\ /a)\ p(y/\alpha\ a)\ p(\alpha\ /a)$$

where the $p(\alpha\ /a)$ are a function of the priors. Of course it is possible to treat *L* as root and set the value of all of the $p(\alpha\ /a)$ to unity. But then indeed *L* is a very strange 'cause', and we are simply left with a representation requiring twice the table entries as a representation of the type illustrated in Figure 12(a).

Thus all discussions to follow are in terms of latent variable models such as the one in (b).

The representational power of the latent variable construction described in the preceding paragraph, and of signal model presented in Section 3 are identical. Identical representation power means that any distribution, $p(x_1, \ldots, x_n, a)$, which can be represented by a causal process, can also be represented by a latent variable and vice versa. First consider constructing a latent variable corresponding to a causal process. By the definition of a causal process we have

$$p(x_1, \ldots, x_n, \sigma_1, \ldots, \sigma_n, a) = \prod_{i=1}^{n} p(x_i/\sigma_i)\ p(\sigma_1, \ldots, \sigma_n/a)\, p(a) \quad (4)$$

By summing out the signals, σ, we have

$$p(x_1, \ldots, x_n, a) = \sum_{\sigma_1} \cdots \sum_{\sigma_n} p(x_1, \ldots, x_n, \sigma_1, \ldots, \sigma_n, a)$$

Constructing the latent variable, *L* , so that it has states $\langle\sigma_1, \ldots, \sigma_n\rangle \in \sigma_1 \times \ldots \times \sigma_n = L$ we can write

$$p(x_1, \ldots, x_n, \langle\sigma_1, \ldots, \sigma_n\rangle, a) = \prod_{i=1}^{n} p(x_i/\langle\sigma_1, \ldots, \sigma_n\rangle)\ p(\langle\sigma_1, \ldots, \sigma_n\rangle/a)\, p(a) \quad (5)$$

and

$$p(x_1, \ldots, x_n, a) = \sum_L p(x_1, \ldots, x_n, \langle \sigma_1, \ldots, \sigma_n \rangle, a)$$

So if we build conditional probability tables for the x_i so that the $p(x_i / \langle \sigma_1, \ldots, \sigma_n \rangle)$ in the latent variable model is equal to $p(x_i / \sigma_i)$ in the causal process model, and set the $p(\langle \sigma_1, \ldots, \sigma_n \rangle / a)$ in the latent variable model equal to the $p(\sigma_1, \ldots, \sigma_n / a)$ in the causal process model, we have built a latent variable model representing the same probability distribution as the causal process model.

Now consider constructing a causal process given a latent variable and its cause. In the most general case, we can simply regard *L* as the lone signal generated by the cause, thus making full state information available to all effects, x_i. Even though this most general case is somewhat trivial, it demonstrates equivalence.

The nature of more interesting transformations from latent causes to signals is evident in the construction above. In this construction, there will be many duplicate valued entries in the conditional probability tables for the effects. Thus given a latent variable and its effects, many nearly identical entries in the conditional probability tables will be evidence for the presence of signals.

Therefore it is possible to re-examine data from which latent causes have been inferred to detect the possible presence of signals.

Thus we have shown that the Communicating Causes model and the latent variable model are representationally equal. In the next section however, we will show that when signals are present, the Communicating Causes model is often a much more economical computational representation. It is more economical both in terms of space and time. We also feel that signal model is more intuitive in its relationship to a modeled domain.

5.2 ADVANTAGES OF EXPLICIT MODELING OF SIGNALS

In this section we will show that whenever correlations among effects arise from correlated signals, directly modeling these signals will be at least exponentially more efficient in time, space, and knowledge acquisition effort than embedding them in latent causes. The basic source of the these savings was already evident equations in (4) and (5) above.

We emphasize that we are only addressing the modeling of correlations among different signals produced by the same causal process and their propagation through the model; we are not claiming to model all possible correlations. Further, we are only claiming that our methods are representationally more efficient than the common correlation modeling technique usually employed to model such correlations.

At this time, we consider the existence of even more efficient modeling methods for signal generated correlations to be an open question. We have shown above that other common methods for modeling correlations, such as the introduction of extra edges or the introduction of latent variables at the same level as the correlation producing causes, are not appropriate.

For a latent cause introduced as an effect of a process generating correlated signals, the key comparison is between the table size of the conditional probability tables for the x_i. In equation (5), the table size required for each x_i is exponentially larger that the table size required by equation (4). From a domain point of view, this is because the latent variable approach supplies each x_i much useless information. The latent variable model notifies each x_i of the value of every signal generated by the cause, even though, in a physical sense, most of this information is of no use. The explicit signal model provides only information about the signal which is actually sensed and interpreted by the x_i.

The above comparison is misleading in one sense. While it is true that the table sizes for the x_i are exponentially larger, these larger tables are only of the same order as the table required to represent the joint distribution of the signals. But this is only true in simple situations such as shown in Figure 12, in which each signal has but a single cause.

We will now present an example to illustrate that savings can be truly exponential. Consider the model shown in Figure 13(a). This model shows a context in which the model of Figure 12 might appear. In this expanded model, *a* is no longer the only cause of *y* ; *b* is also a cause.

Suppose that from domain knowledge, we believe that *a*'s mechanism for causing *y* is generation of the signal ; *b*'s mechanism is the same. Process *y* can sense signal , but has no way of discerning which process caused this signal. Further we believe that *a* generates signals α and so that they are correlated, and that *b* similarly generates and γ. This knowledge is diagrammed in Figure 13[10].

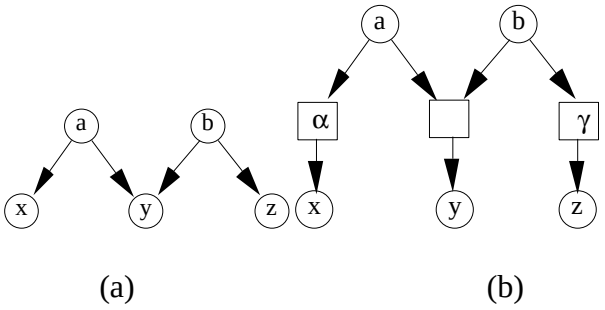


Figure 13: Multiple Causes of a Signal

Now compare the complexity of the full models underlying the diagrams of Figure 13. The comparative complexity is illustrated in Figure 14.

We will focus our comparison on the size of the conditional probability tables required by event *y*[11]. The table for *y* will contain $\times$ y under the assumptions of Figure 14(a). Under the

[10] This may initially appear to be a *very* special case. But it is our hypothesis that this is often the case, but has gone unrecognized and has been "covered up," if you will, by the use of latent variables.

[11] The table size for the latent variable itself exactly matches the table size to represent the correlations of the signals.

assumptions of Figure 14(b) the table will contain $\alpha \times \quad \times \quad \times \gamma \times y$ entries.

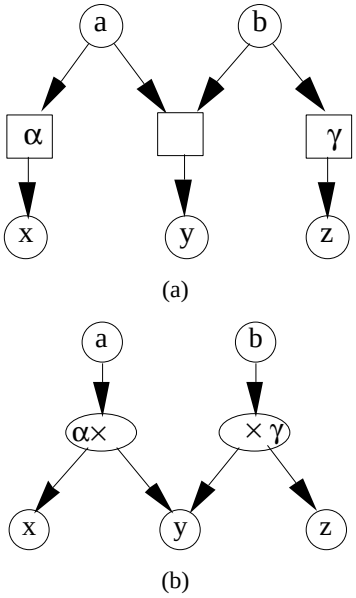


Figure 14: Comparative Complexity

This huge advantage for the Communicating Causes model may be partially offset however if a table is (and it may not be, as will be discussed below) required for . In this example, the (possibly required) table's size is $a \times b \times$. Thus when a table is required, the ratio of size of tables required by the Communicating Causes model to that of the latent variable model is given by

$$\frac{(a\ b + y)}{y\ \alpha\ ^{2}\gamma}$$

Thus if all variables were binary the size of the required tables for the Communicating Causes model would be $^{12}/_{32}$ of the size of the latent variable.

But often in the Communicating Causes model, a function will suffice for instead of a table. This function (or table) actually specifies how signals combine at a receiver. This is a function which is determined by the domain, and might be as simple as a Fuzzy OR, or as the RMS of the received power (if the signals are radio signals). In any case the function describes how the signals *physically* combine.

In a case where a function suffices, the Communicating Causes model will require only $^{1}/_{\alpha\gamma}$ of the storage required by the latent variable whenever the signal combination function can be exploited in the table or $^{1}/_{\alpha\ \gamma}$ when it cannot be.

The above example is suggestive of the representational efficiency achievable by use of the Communicating Causes approach. It of course does not *prove* that comparable savings can be achieved in the general case. The general case remains an open question.

5.3 COMPUTATIONAL METHODS FOR EXPLICIT SIGNAL MODELING

It is clear that traditional algorithms for prediction and inference in causal models are not applicable to models with explicit representation of correlated signals. But there are other computational procedures already reported in the literature which are adaptable to such models. Traditional algorithms such as those in Hugin, are not applicable because they have no way of representing the signal generation probabilities, $p(\sigma_1, \ldots, \sigma_n / a)$, appearing in equation (4). Since such probabilities have the potential for propagating conditional dependencies throughout a causal model, Hugin-like algorithms are not readily adaptable because of growing clique sizes. Techniques readily adaptable to explicit signal modeling already exist however. Sampling and simulation methods such a those surveyed in [Fung 93] are all readily adaptable to models with explicit signal representation. The technique in [Fung 93] seems especially appropriate for such models.

While it may seem very unaesthetic that explicit representation of signals eliminates the practicality of known closed form algorithms, it is worth noting that many widely used commercial systems successfully use simulation and sampling based techniques. Among the practical advantages of such techniques are their "any time" character. Now the advantages of more compact modeling of relevant correlations can be added to their other advantages.

6. CONCLUSIONS AND FUTURE WORK

We have introduced a physically based model of causal influence, the Communicating Causes model, which we feel provides a powerful reason to doubt that conditional independence can generally be inferred from causal relations, i.e. a powerful reason to doubt the applicability of the Causal Markov assumption. Our model, if it is correct, shows that such conditional independence, when it exists at all, must arise from the nature of each particular causal *process* rather than from causality itself. We have shown that any empirical evidence purporting to support the truth of the Condition must be considered suspect whenever latent variables have been "discovered."

The appropriateness and value of our model is empirically testable and should be tested. At one level, domain experts can attempt to identify signals as physical phenomena important in the operation of their domain. At another level, existing models which have incorporated latent variables can be examined for evidence of the presence of signals: Section 5.2 suggests that many near duplicate values in conditional probability tables may be a sign of the presence of signal phenomena in the domain.

We feel that knowledge engineers should no longer feel free to assume conditional independence based on the existence of causal influences alone. We feel that they must look more deeply into the actual causal mechanism of each causal process before assuming conditional independence. In particular, we recommend that they try to identify the mechanism by which causal influence is achieved before invoking conditional independence.

We have provided Knowledge Engineers a criterion by which to determine if their model is complete in the sense that modeled causes need not be further decomposed.

We have shown that if correlations among effects are present, that it is not appropriate to model these correlations by adding additional edges to the model whenever it is suspected that the correlations might arise from communicating causal processes.

We have shown, independent of the validity of the Condition, that when signals are adopted as the mechanism of causal influence, explicit modeling of these signals can provide great computation efficiency.

In the future we will attempt to work out the implications of the Communicating Causes model to data mining, etc. We have already begun an empirical investigation of the magnitude of prediction and inference errors which can be introduced by inappropriate use of the Condition.

Acknowledgments

Thanks to Greg Cooper, and especially to Marek Drudzel, for their comments and suggestions on early drafts of this paper. Special thanks to Paul Losiewicz without whose advice and encouragement this paper would not have been completed.

References

Elby, A. Should we explain EPR correlations causally?. *Philosophy of Science*, v. 59 pp.16-25 (1992)

Fung, Robert and Del Favero, Brendan. Backward Simulation in Bayesian Networks. *Uncertainty in Artificial Intelligence, Proceedings of the Tenth Conference.* San Mateo, CA: Morgan Kaufmann (1994)

Lemmer, John (1993) Causal Modeling. *Uncertainty in Artificial Intelligence, Proceedings of the Ninth Conference.* San Mateo, CA: Morgan Kaufmann (1993)

Pearl, Judea. *Probabilistic Reasoning in Intelligent Systems.* San Mateo, CA: Morgan Kaufmann (1988)

Spirtes, Peter, *et al*. *Causation, Prediction, and Search*, Springer-Verlag (1993)