Available online at www.sciencedirect.com
ScienceDirect
International Journal of Approximate Reasoning
46 (2007) 257–273
INTERNATIONAL JOURNAL OF APPROXIMATE REASONING
www.elsevier.com/locate/ijar

Does non-correlation imply non-causation?

Eric Neufeld *, Sonje Kristtorn

Department of Computer Science, 110 Science Place, Saskatoon, SK, Canada S7N 5C9

Received 29 June 2005; received in revised form 14 March 2006; accepted 21 September 2006
Available online 30 November 2006

Abstract

The Markov condition describes the conditional independence relations present in a causal model that are consequent to its graphical structure, whereas the faithfulness assumption presumes that there are no other independencies in the model. Cartwright argues that causal inference methods have limited applicability because the Markov condition cannot always be applied to domains, and gives an example of its incorrect application. Cartwright also argues that both humans and Nature, fairly commonly, design objects that violate the faithfulness assumption. Because both arguments suggest that data is not likely to be ideal, we suggest that problems of the theory be separated from problems of the data. As regards the Markov condition, conflicted intuitions about conditional independence relationships in certain complex domains can be explained in terms of measurement and of proxy selection. As regards faithfulness, we show that violations of this assumption do not affect the predictive powers of causal models. More generally, the criticisms of causal models, taken constructively, reveal the subtlety of the ideal, while clarifying the source of problems in data.

1. Introduction

Causal graphs have been studied as such for more than a decade. Originally introduced as Bayesian nets [14], they demonstrated the practicality of purely probabilistic reasoning to an AI community that believed probability theory was epistemologically inadequate

* Corresponding author.
E-mail address: eric@cs.usask.ca (E. Neufeld).

0888-613X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2006.09.013

even for mundane knowledge. Causal graphs gave both a compact representation of joint distributions of many variables, and sound and efficient inference algorithms.

Earlier, many groups sought alternatives to probability for the purpose of representing uncertain knowledge. Several groups pursued non-numeric, or symbolic alternatives, such as endorsements or the various default logics. (See [9] for discussions.) The reasons for different strategies varied, but it would be fair to say that many believed that intelligent agents were capable of sophisticated reasoning strategies without numeric information, and furthermore, that accurate statistics were rarely available. In medical diagnosis, diseases can evolve so quickly that it is difficult to collect accurate statistics before the population changes significantly.

Other groups, citing limitations of the expressive power of traditional probability, pursued alternate numeric calculi. Three prominent formalisms at that time were Certainty Factors (the calculus driving expert systems like MYCIN), Fuzzy Logic, and belief functions (also known as Dempster–Shafer theory). Representatives of these formalisms formed the core of the early Uncertainty in Artificial Intelligence (UAI) community [9], and early papers on causal graphs also appeared at this time. The result was a thorough discussion of the foundations of uncertain reasoning.

Initially, the 'causal' aspect of causal graphs was informal. In simple settings such as diagnostic bipartite causal graphs (e.g., the work of Peng and Reggia [17] translated to causal graphs), this was abundantly clear. Root nodes were unconditionally independent diseases, leaf nodes were conditionally independent symptoms, and arrows pointed in the direction of causality. Causal graphs were found to have many useful qualitative interpretations. Pearl and Verma [15] offered probabilistic definitions of potential and genuine causality that not only gave philosophical justification for this phenomenon, but also prescribed the inductive causation (IC) algorithm for recovering causal structure from sufficiently rich numeric data. The algorithm could also identify certain spurious associations, that is, variables whose correlation was due to both variables being caused by a hidden unmeasured variable. Verma and Pearl [25] gave another algorithm.

Spirtes, Glymour and Scheines contemporaneous developed algorithms to induce models from data, taking into account from the outset the problems of noisy data. Spirtes et al. [22] proposed the PC algorithm for inferring causal structure. This algorithm and several others are discussed in [23]. As well, Druzdzel and Simon [6] give a different algorithm that presupposes identification of exogenous variables.

The present discussion is based on the generalized algorithm in [16]. The soundness of Pearl's inductive causation algorithm rests on two important assumptions. The first is that correlation implies the presence of causality somewhere. More specifically, if A and B are correlated, then either A causes B, B causes A, or both A and B have a hidden common cause. The second is the faithfulness [23] assumption (also known as stability [16]), which, "conveys the assumption that all the independencies embedded in [the probability distribution] P are stable, that is, they are entailed by the structure of the model D and remain invariant to any change in the parameters Θ_D" [16]. The parameters are the numerical conditional probabilities stored at each node, and the number of oriented arcs in the output depend on the presence of conditional dependencies and independencies satisfying the Markov condition.

Pearl's algorithm also assumes ideal data, that variables can be measured directly (i.e., no need for proxies), that variables have relatively few direct causes and that all distributions are error-free. Such assumptions rightly concern working statisticians, who routinely

deal with far more troublesome data. Regardless, we believe that the idealized conceptual framework gives causal models their power, and in the sequel, we respond to criticisms of both the Markov condition and the faithfulness assumption, and show how the formalism itself can be used to troubleshoot models containing counter-examples.

2. Notation and terminology

A causal graph is a pair (D, P), where $D = (V, E)$ is a directed graph, V is a set of variables and E is a set of directed edges (arcs) between variables. p is a probability distribution over the variables in V. For any variable A, parents(A) denotes the parents of A, the direct predecessors of A, or the direct causes of A. Associated with A is a local distribution $f(A, \text{parents}(A))$, which gives a distribution for A for any set of values that the parents(A) take on. Moreover, the distribution p can be decomposed into these independent conditional distributions at the nodes. Commonly f is the familiar discrete conditional probability distribution. For the present, it suffices to say that the formalism generalizes beyond discrete and Gaussian variables (in the form of structure equations, or path analysis), but for this discussion, we use discrete distributions in the text and a Gaussian example in images.

For discrete variables, the graph's structure encodes the information that P factors into the product of the conditional distributions stored at the nodes. That is,

$$P(v_0, \ldots, v_n) = \prod_i p(v_i|\text{parents}(v_i)). \tag{1}$$

Variables A, B are conditionally independent given a set of variables C if

$$p(A, B|C) = p(A|C) \cdot p(B|C)$$

for all outcomes of A, B and C. If C is empty, then A, B are unconditionally independent. For continuous variables, the corresponding concepts are correlation, covariance and partial correlation, and *association* is used as an umbrella term. The factorization in Eq. 1 has the property that vertices obey the *Markov Condition*, that a vertex is conditionally independent of its non-descendants given its parents. In the setting of a causal graph, the Markov condition implies many other conditional independencies. These can be detected from D alone using a graph-theoretic criterion called *d-separation* [16].

Definition 1 (Potential Cause). [15] A is a potential cause of C if there is a variable B and a context (set of variables) S such that

1. A, B are independent given S,
2. there is no set D such that A, C are conditionally independent given D, and
3. B, C are dependent.

A simplified version of Pearl's causal graph construction (IC algorithm) follows:

1. For each pair of vertices A, B in V, search for a subset S of V (including the empty set) such that A is conditionally independent of B given S. If no such S exists, add an undirected edge between A and B.
2. For each collinear triple of vertices A–C–B, where A and B are non-adjacent, test whether A, B and C satisfy the relationship of potential cause as in Definition1. (That

is, simultaneously test that both A and B potentially cause C.) If yes, add head-to-head arrows at C. Repeat this step until no more arrows can be added. (It is acceptable to obtain double-ended arrows, but this is not discussed here.)

3. For all remaining undirected arcs, add arrows, but do not create any new structures like those in Step 2, and do not create any directed cycles. Given these constraints, certain arcs may be oriented in either direction. Such arcs should be left undirected.

The output of this algorithm is a graph containing undirected arcs, directed arcs (arrows), and arcs directed in both directions. Directed arcs indicate causal links. Arcs directed at both ends indicate the location of hidden causal variables. Undirected arcs indicate insufficient information to orient an arc.

When constructing the directed causal graph from data, this algorithm twice makes critical use of the faithfulness assumption. In Step 1, it places an arc between two nodes only if no set S makes them independent. Under the usual interpretation of faithfulness, the absence of any form of independence forces a link to be added, since the only dependencies are those implied by structure.

We argue that the intuition behind faithfulness is the idea that causation implies correlation. Under this interpretation, we think of the IC algorithm in terms of the complementary action of not adding a link when independence is discovered. All links added to the graph have the potential of being oriented to imply causation. Thus, independence (or the lack of dependence, or non-correlation) in any setting means no causal link is added, or, non-correlation implies non-causation.

The faithfulness assumption is also used in Step 2, embedded in the definition of potential cause. Under the usual interpretation of faithfulness, consider all possible orientations of the arcs connecting A, B, C as shown in Fig. 1.

Using the faithfulness assumption in the traditional sense, independencies in the data are implied by structure only, and the head-to-head structure is the only one of the four that implies the data.

Under our interpretation, the assumption that causation implies correlation rules out the first three possibilities, by reasoning as follows. Consider a setting where A and B are independent, and there is a path from A to B. If causation implies correlation, the first two orientations are easily ruled out, since A and B are not correlated. In the third case, one also expects correlations between effects of a common cause. However, it is straightforward to use the factorization properties of a causal graph to verify that only the last graph is always consistent with the definition of potential cause, regardless of assignment of conditional probability distributions to the vertices. That is, the definition of potential cause is based on the assumption that causation implies correlation.

A causal inference algorithm such as this one cannot do much without an assumption like this, although it is easy to construct counter-examples that show the faithfulness assumption to be imperfect. Consider a graph such as that in Fig. 2a.

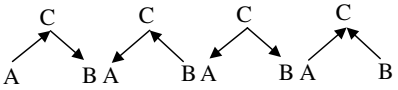

Fig. 1. All possible arc orientations from A to B.

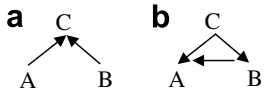

Fig. 2. Two causal models of A, B and C.

Because small (3 node) graphs with no unconditional independence can have three different labelings (that is, the first three labelings in Fig. 1), the graph in Fig. 2a is the smallest graph that the IC algorithm can orient. However, it is straightforward to construct a graph with the topology of Fig. 2b: just use the graph in Fig. 2a to compute the conditional probability distributions for each node in the second graph. Although this is a simple exercise, it is interesting for two reasons: (1) it introduces an arc where none existed before, and (2) the causal directions of the original arcs in Fig. 2a are reversed.

Note that the algorithm does not impose the Markov assumption on its model. The independencies used to construct the graph come from the data.

3. Cartwright's criticism of the Markov assumption

Cartwright [3] considers the Markov condition under two aspects: The first of these concerns temporal relationships between causes. Causes do not operate 'across temporal gaps' (p. 107), so conditioning on the parents of a variable makes it independent of the rest of its ancestor variables. The second aspect affirms the mutual independence, or 'screening off', of variables, given their common parents. Thus, conditioning on the parents 'screens' their shared effects from each other. It is this 'screening off' aspect of the Markov condition that is the subject of her argument.

The SGS position [23], according to Cartwright, is that the Markov condition holds for "causal graphs of deterministic or pseudo-deterministic systems in which the exogenous variables are independently distributed" [23, p. 57]. She describes a pseudo-deterministic graph as a (correct) indeterministic subgraph of a more complete (and likewise correct) deterministic graph – bearing in mind that correctness, not completeness, is the achievable goal. Glymour describes a pseudo-indeterministic system more precisely as a system without feedback obtained from a larger system by "marginalizing out some exogenous causes" [8]. In any case, that the Markov condition holds for deterministic systems as well as for apparently indeterministic subsystems of deterministic systems is, she says, "trivially true".

When the relationship between cause and effect is deterministic, causes fix the values of their effects. So, for example, in the simple case of a single cause of two effects, if I know the value of the cause, I know the value of each effect. The effects are independent of each other, given the cause, since knowing the value of one effect tells me nothing about the value of the other that I do not already know, if I know the value of the cause. Formally, $p(E_1|C, E_2) = p(E_1|C)$, where C represents the cause and E_1 and E_2 represent the effects of C. This means that the joint probability of two effects, given the cause(s) of those effects, will factor: $p(E_1, E_2|C) = p(E_1|C)p(E_2|C)$. However, according to Cartwright, deterministic causality is uncommon and the Markov condition does not (in general) hold where causality is probabilistic.

When a cause operates probabilistically the relationship between the value of the cause and the value of the effect is probabilistic: for a given value of the cause, there is a

probability distribution over the possible values of the effects. (In the deterministic case, there is one possible value for each effect, given the cause, with probability 1.0; any other values have probability 0.0.) Cartwright argues that the value of one effect of a probabilistic cause will almost always provide some information about the value of a second effect, even if I know the value of the cause. In other words, there is no screening off here, no conditional independence. Likewise, the joint (conditional) probabilities will not factor.

Why should the fact that relationships between cause and effects are probabilistic lead us to draw this conclusion? Here Cartwright points to interactions between the operations of a cause to produce its different effects, something that is not directly represented in causal graphs. In the case of deterministic causality, these interactions need not be considered since a cause infallibly produces all of its effects. But in the case of probabilistic causality, we have to consider not only the probabilistic relationships between a cause and each of its effects but also the relationships, likewise probabilistic, between those causal relationships. When a cause can occur and yet not produce one or more of its effects, such interactions have to be taken into account when calculating the probability of one or another effect, given the cause. Cartwright apparently regards the screening-off aspect of the Markov condition as an attempt to evade or at least resolve this complexity by positing a "'split-brain' model of the common cause" [3, p. 108], according to which a cause operates independently to produce each of its effects. If the causal operations of a single cause are independent of each other, we need not consider the relationships between them. The heart of her objection seems to be an assertion that this simply is not so, or is very rarely so; usually it is a matter of 'joint operations', that is, there is a (probabilistic) relationship between the operations of a (probabilistic) cause to produce its various effects. So, even when we know the value of the cause, knowing the value of one effect is likely to tell us something about the value of the other.

Cartwright concedes the divorce-age-candy example of Blalock as an example that is consistent with the split-brain model. In that example, eating candy and divorce become conditionally independent, conditional on age; otherwise they have a negative association. The aging mechanism is complex and it seems intuitively agreeable that change in fondness for candy is governed by a mechanism independent of that which causes divorce – especially when compared to the smaller world of colds causing fevers and stuffy noses, where it seems almost impossible to imagine that the mechanism by which a cold causes one of these symptoms is independent of the other. However, given some reflection, it is not hard to imagine a few other variables that are comparable to age in complexity – for example, income, gender, nationality, personality type, and height. Each variable reflects an accumulation of inputs from a huge number of variables and to some extent hides a good deal of direct causal structure. As well, the outputs from the many variables become distilled into a single link. Many features of age are highly correlated – perhaps Blalock's example is just one where many things balance out.

It is in this context that Cartwright presents her factory example. The main lines of this argument are as follows: There are two factories, D and P. Both produce a certain chemical, C, that is used immediately in a sewage plant. Factory D produces the chemical with probability 1.0, or deterministically. Factory P, on the other hand, produces the same chemical with probability 0.8, that is, probabilistically. Moreover, whenever P produces the chemical, it also produces a pollutant, B, as a byproduct. However, the owner of factory P maintains that the pollutant is produced when the chemical is used in the sewage treatment plant and, in support of this claim, advances an argument that assumes the

Markov condition in its screening-off aspect, as follows: if the pollutant were a byproduct of factory P, then the probability of the pollutant would be independent of the probability of the chemical, given the factory that produced the chemical. Factory P argues that if it were responsible for the pollutant then $p(C,B|P) = p(C|P)p(B|P)$. But in fact $p(C,B|P)$ is 0.8 while the two factors are likewise each 0.8; therefore, argues factory P, it is not responsible for the noxious byproduct, since $0.8 \neq 0.8 \cdot 0.8$.

But factory P's argument is wrong, says Cartwright. Factory P is responsible for the byproduct. Factory P's error lies in its screening-off assumption. Because factory P produces both the chemical and the byproduct probabilistically, the probabilities of C and B are not independent given P and hence do not factor, that is, $p(C,B|P) \neq p(C|P)p(B|P)$; knowing something about the presence of the byproduct tells us something about the presence of the chemical. To pursue the 'split-brain' metaphor, the mechanisms whereby factory produces B and C are not independent.

In this example a cause, factory P, has two operations that produce two effects: one operation produces the chemical while the other produces the pollutant. The pollutant is a produced, apparently with probability 1.0, as a byproduct of the process that produces the chemical. Because of this relationship, the effects are conditionally dependent: even if we know the value for factory P, knowing the value of, for example, the chemical variable provides information about the value of the byproduct variable, B, that is not derivable solely from the value of their common cause, P. The value of the cause variable does not tell us whether the cause has been effective, whether it has, as Cartwright puts it, 'fired' and produced the chemical; it tells us only the probability that this is so. Knowing the value of the byproduct variable, B, provides this information and thereby provides information about the value of the chemical variable, C, that we would not know simply by knowing the value for P.

4. Response to Cartwright's criticism of the Markov condition

This example raises a few questions. First, there is a question about the consistency of the distribution as described. It would appear from the example that $p(C|D,P) = 1.0$, $p(C|\overline{D},P) = 0.8$, $p(C|D,\overline{P}) = 1.0$, and $p(C|\overline{D},\overline{P}) = 0.0$. Using Jeffrey's rule, and a little algebra,

$$
\begin{aligned}
p(C|P) &= \frac{p(C,P)}{p(P)} \quad \text{(Definition of Conditional Probability)} \\
&= \frac{p(C,D,P) + p(C,\overline{D},P)}{p(P)} \quad \text{(Jeffrey's Rule)} \\
&= \frac{p(C|D,P)p(D|P)p(P)}{p(P)} + \frac{p(C|\overline{D},P)p(\overline{D}|P)p(P)}{p(P)} \quad \text{(Product Rule)} \\
&= p(C|D,P)p(D|P) + p(C|\overline{D},P)p(\overline{D}|P) = 1.0p(D|P) + 0.8p(\overline{D}|P) \\
&= p(D|P) + 0.8(1.0 - p(D|P)) = p(D|P) + 0.8 - 0.8p(D|P) \\
&= 0.8 + p(D|P)(1.0 - 0.8) = 0.8 + 0.2p(D|P).
\end{aligned}
$$

Hence, if $p(C|\overline{D},P) = 0.8$ then $p(C|P)$ must have a value greater than the 0.8 given in the example. These are equivalent only if $p(D|P) = 0.0$; that, however, would mean that there

was a dependency between D and P, which is contrary to the graph for the example [4]. We could contrive the distribution so that $p(C|P)$ comes out to 0.8, using the above calculation. This would require that a value be assigned to $p(D|P)$ or, effectively, to $p(D)$ since, according to the graph, D and P are unconditionally independent. It would also require the assignment of an appropriate value (less than 0.8) to $p(C|\overline{D}, P)$. However, the same value would have to be assigned to $p(B|D, P)$ and, equivalently in this case, to $p(B|P)$, contrary to the example. Although we consider this a flaw in the presentation of the problem, it is a relatively minor issue. The next two questions go deeper.

A second question is related to an objection raised by Glymour [8] in a response to Cartwright. Glymour claims that it is very difficult if not impossible to find an example of a system that violates the Markov condition, *if all causes are considered*. We observed earlier that a pseudo-indeterministic system, where exogenous causes are 'marginalized out', does not violate the Markov condition. But apparent counter-examples, wherein the Markov condition does not hold, are possible if causes are left out in other ways. For example, common causes may have been overlooked. Thus, an apparent counter-example may overlook the fact that variable values have been collapsed into a reduced set or it may involve a mixture of different systems. Moreover, there are physical systems where state descriptions fail to screen off prior state descriptions but this violation of what Cartwright describes as the first aspect of the Markov condition is, according to Glymour, generally taken to indicate some incompleteness in the state descriptions. In fact, Glymour maintains that most, if not all, counter-examples derive from ignorance of one kind or another.

It would seem that examples like the factory example might be vulnerable to an objection of this sort. If, for example, a common cause of a parent cause and one of its effects is omitted from consideration, the effects will not be independent, given their common parent. But this is because a common cause has been omitted, not because the Markov condition does not hold for the causal system the example intends to describe. The Markov condition does hold – so this objection goes – if the incompleteness is remedied. If the common cause is included, the two effects will be independent, once we condition on all the parents.

We have two responses to the above. On the one hand, we can interpret Cartwright's claim about the inapplicability of the Markov condition as implying that the Markov condition is an impossible ideal, and Glymour's response as, in some sense, confirming this. One can strike a philosophical compromise by saying that the unachievable ideal is nonetheless a useful tool, analogous to propositional logic. By analogy, very little, if any, deep knowledge can be captured succinctly in propositional logic. Regardless, propositional approximations of reality are widely used. The second response is to look seriously at the nature of what we call variables, and the ways we devise to measure them.

The philosophical compromise already exists in the way science is done, i.e., we use logic in the small and large roughly as follows: From experiences, we generalize to logical sentences, which we use until new experiences provide counter-examples. At that point, we have two choices. One, we can revise or reject the theory. Two, we can reject the evidence. Perhaps we have been fooled with false data or our eyes have been tricked, or we become convinced that the original evidence was a fluke. This is reasonably common in medical diagnosis – a physician may well order tests to be redone before rejecting a reasonably strong diagnosis [18].

Alternatively, we can explain Cartwright's objection in terms of measurement. Measurement is that elusive idea of how we classify and measure the world [10]. At the level

of discrete measurement ('natural kinds'), a theory of measurement tells us how to decide when the proposition 'this is a stone' is true, and when it is false, so that we can count how many stones we have. This information is not perfect: over the long run, we will make counting errors with some probability. We may decide to weigh the stones, to distinguish stones from rocks and pebbles, but then we have a new problem – defining an apparatus that weighs (or measures mass of) objects. The conceptualization and implementation of this apparatus is also a process that is subject to dispute.

The factory problem can be modeled as a measurement problem. How do we measure such events as P and D? A factory is not like a circuit with a switch, a battery and a light. Machines must be warmed up, and cooled down. The consequences of certain effluents may take considerable time to obtain. The factory may operate at different levels of production during different seasons and measuring instruments may behave differently at different times. Achieving universal agreement on whether or not the factory is 'running' on any day is impossible. Of course, a set of reasonable observers will likely come up with *similar,* but not identical definitions. Any particular definition will provide an approximation, but none will agree exactly.

In applied statistics, researchers choose proxies for this type of variable. Suppose you wish to measure *Quality of Computer Science Graduate Program* for a set of Canadian universities. (A similar variable exists in a study by Rodgers and Maranto [20].) There is no way to measure this directly, but a possible proxy might be the number of Computer Science graduate students with NSERC scholarships, using the reasoning that NSERC scholarships are the most competitive and go to the brightest applicants, and the brightest applicants will chose the best graduate programs. However, some universities might reply that many students choose universities for geographical reasons, and moreover, NSERC scholarships are not awarded to international students. One could then consider counting the number of faculty with external research grants. Some universities argue that they have mostly junior members who are not well established and only a few senior members with large grants. This leads to the possibility of measuring the total value of all research grants. In the end, a combination of several variables – class size, perceived reputation – form a figure of merit that facilitates comparison. It is generally difficult to get consensus on any particular figure.

Although it is possible to model Cartwright's factories as light bulbs on a simple circuit, it is more realistic to imagine a more complex model that might make selection of measurement method difficult. So, for example, we can select a certain threshold value on a certain output valve, or we can analyze water supplies. Consider the model depicted in Fig. 3. M_1 is one possible measurement taken to determine whether the factory is running. M_2 is a different measurement. Moreover, M_2 is a common cause of both M_1 and B. Both M_1 and M_2 are proxies for whether the factory is running, and one is much better than the other. The relevant conditional probabilities are as follows:

$$p(M_2) = 0.1$$
$$p(M_1) = 0.15$$
$$p(M_1|M_2) = 0.6$$
$$p(M_1|\overline{M}_2) = 0.1$$
$$p(M_1) = 0.15$$
$$p(C|M_1) = 0.8$$

$$p(C|M_2) = 0.72$$
$$p(C|\overline{M}_1) = 0.6$$
$$p(B|M_1, M_2) = 0.8$$
$$p(B|M_1, \overline{M}_2) = 0.8$$
$$p(B|\overline{M}_1, M_2) = 0.6$$
$$p(B|M_1) = 0.8$$
$$p(B|M_2) = 0.72$$
$$p(C, B|M_2) = 0.528$$

These probabilities were obtained using a simple brute force search, constrained to report only distributions where C and B are independent given M_1 and, dependent given M_2. We took care to select a distribution where M_1 and M_2 were otherwise as similar as possible. Both co-occur frequently, and both present C and B with approximately similar frequencies. We also took care to find distributions similar to the one in Cartwright's example.

From the topology of the graph, it is easy to show that if we condition on $M1$, the two factory outputs are independent, and if we condition on M_2 they are dependent. The example further shows that there is no hidden independency buried in the conditional probability distributions. Assuming that both measurements would be acceptable to a human observer as reasonable proxies, one implies the conditional independence relationship associated with the Markov condition, and the other implies the dependence shown in Cartwright's example.

This gives us a different perspective on Cartwright's argument. Glymour's claim that the Markov condition always holds given all causes are represented is as difficult to defend as Cartwright's claim that it almost never holds. But given that the Markov condition can hold, and the problem is one of choosing the correct metric, then the problem of finding a measurement consistent with the ideal we are trying to measure is a familiar one. Deciding the 'true' relationship between C and B may be difficult if disputes over methods of measurement cannot be resolved. However, this shows that whether the Markov condition holds or not in the ideal the proxy was intended to represent is an empirical, rather than a theoretical, question.

Such an argument may seem strange in the case of a fictional example. Someone who wants to dispute the example can perhaps claim that a crucial cause has been overlooked. But in the case of a fictional example, unless the alleged missing cause is implied in the example in the first place, it is difficult to see how such an argument can be sustained

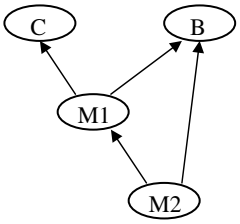

Fig. 3. A causal model for the measurement example.

without begging the question. On the other hand, the one who puts forth the example can stipulate that no such cause has been overlooked. Thus, Cartwright's factory example specifically states that all causes of the chemical and the byproduct are included. Again, there seems to be an element of question-begging here. And therein lies another difficulty, also raised by Glymour: the factory example is entirely fictional yet it is intended to refute what Glymour calls a 'hypothesis about nature' [8]. Glymour argues that we can always imagine a counter-example to a hypothesis about nature but such imagined counter-examples cannot refute the hypothesis.

Glymour makes a similar rejoinder to another argument Cartwright advances. This argument, which might be considered a negative conceptual argument, is embedded in a kind of 'what are the odds?' argument: How likely is it, when we have a joint probability distribution over the effects of a common cause, that the distribution will be such that the effects are conditionally independent? The answer, she says, is "not very" and consequently screening off cannot be assumed; there must be good evidence for it in a particular case. This line of argument is buttressed by the claim that there is nothing in the concept of causality that implies 'screening off', that "nothing in the concept of causality, nor of probabilistic causality, constrains how nature must proceed" (*The Dappled World*, p. 109). Glymour responds that something may be true in nature even though it is not a part of the relevant concepts. He also argues that it is entirely possible that the way human beings think about causes may indeed be inseparable from the Markov condition, whether this derives from hard-wiring or experience and whether this is explicit in our concept of causality or no. In fact, Glymour regards the Markov condition as a kind of 'rule of thumb', an empirical heuristic that is almost always valid and gives good results, rather than as a component of a well-established scientific theory on the one hand or an a priori principle on the other.

Thirdly, there is the question of the role Cartwright ascribes to the Markov condition. She argues that, given the situation described in her 'factory argument', B and C must be conditionally independent, given P, if the Markov condition holds in this case. This intuition corresponds to a graph of the situation, depicted in the *Synthese* paper though not in *The Dappled World*, wherein P is the sole cause of both B and C. The Markov condition as applied to this graph implies the conditional independence of B and C, given P. She then points out that the expected conditional independence relationship does not hold, given the distribution in the example. Above, we suggested that whether or not this kind of intuition of conditional independence turns out to be valid can turn on measurement issues. However, measurement issues aside, an expectation of conditional independence can be falsified, as in this example, by the distribution. In the factory example, factory P's argument assumes not only the Markov condition but the graph to which it is applied. However, if B and C are not, according to the distribution, conditionally independent given P, the 'true graph' (as opposed to the intuitive model) will have an arc between B and C and therefore the Markov condition, in its 'screening off' aspect, will not, when applied to this graph, imply the conditional independence of B and C, given P. Any graph produced by the inductive causation algorithm will have an arc between B and C, if this conditional independence relationship does not apply. It is difficult to see, then, that the factory example casts doubt on 'screening off'. More broadly, it is difficult to see how what she has to say about probabilistic causality bears on this issue since a graph generated by the algorithm will show 'screening off' relationships only where the corresponding conditional independence exists. Lemmer [11] gives a similar structure, and, states, paraphrasing with

our variable names and emphasis added, that *assuming* $p(C, B|P) = p(C|P)p(B|P)$ unduly restricts our intuition of what it means for P to be a cause.

The same kind of question arises in connection with the 'what are the odds' argument. It is difficult to see the significance of 'the odds' of conditional independence for the question of the validity of 'screening off' since a graph will display this kind of relationship only where the corresponding independence relationships are found in the distribution. The likelihood of such an occurrence does not seem to matter.

5. Cartwright's critique of faithfulness

A counter-example to a premise is generally certain death for a theory. A single counter-example (and there are infinitely many in the present case) shows that we cannot say that causation implies correlation (in the usual sense of the first order logic), which we have argued above is the intuition of the faithfulness assumption.

However, it is possible to argue that probability of correlation given causation has measure one. For two variables A and C, for any distribution of A, there is exactly one joint distribution of A, C such that the two variables are independent, but in all the remaining uncountably many distributions, the two variables are dependent.

This ingenious Bayesian argument is used for the faithfulness assumption, and assumes that all probability distributions are equally likely. This latter assumption runs into difficulty, clearly expressed by Cartwright [3]. She states

> It is not uncommon for advocates of DAG-techniques to argue that cases of cancellation will be extremely rare, rare enough to count as non-existent. That seems to me unlikely, both in the engineered devices that are sometimes used to illustrate the techniques and in the socio-economic and medical cases to which we hope to apply the techniques. For these are cases where means are adjusted to ends and where unwanted side effects tend to be eliminated wherever possibly, either by following an explicit plan or by less systematic fiddling.

Elsewhere [2], she goes into considerably more detail. In essence, she states that Pearl's argument puts "structure first", and parameters second. However, she claims that one cannot have one without the other, and gives a convincing example. Birth-control pills may cause thrombosis, and thus we try to weaken the strength with which they so do. Thus, she concludes "Getting the cancellation that stability/faithfulness prohibits is important to us". More generally, she argues, that probability and causal structures constrain each other. If the probabilities are fixed, then we are constrained from building certain causal structures, or, (in the case of faithfulness), vice-versa.

It may be easier in applied science and/or engineering to design a counterprocess to cancel certain effects of a process than it is to eliminate the process causing the effect in the first place. Thus, a case can be made that Nature, as engineer, frequently uses the same ploy. Shipley [21] gives examples from nature (photosynthesis) where processes act as counterweights to establish a set point. Since the theory of causal graphs cannot get far without the faithfulness assumption, this is potentially devastating.

This provides an unusual meta-interpretation for Fig. 2a. Let C be skin pigmentation, B be sunscreen usage and A be incidence of melanoma. Perhaps sunscreen usage and melanoma independently conspire to eliminate the expression of skin pigmentation's causal power through probabilistic dependencies and independencies. This seems implausible.

However, in a world where we do not know the true number of causal influences on any effect, it seems possible, if experimental error is considered, that some pair of them might almost exactly cancel, and that we might have unfortunately picked that pair.

6. Whether causation implies correlation

An assumption that precedes the faithfulness assumption is an assumption regarding the existence of causality. Karl Pearson saw cause and effect as an old and simplistic concept, as extreme cases on a continuum of dependence, rendered obsolete by the new ideas of association. Cartwright's view suggests that the new theories of causation are too simplistic and accident-prone to be trusted.

This is our view: The world we inhabit cannot be experienced directly. It is knowable only through our senses, which vary widely among individuals, and measurable only through devices constructed in that same world that are subject to some error. However, simple direct cause and effect relationships may properly exist in this world, and may be measurable in some sense, but the outdegree and indegree of every node may be prohibitively high. If causation does not properly exist in the real world, we could, as in other mathematical sciences, use causality as an idealization from which to make predictions about interventions.

In this ideal world, causality exists. In this ideal world, causality implies correlation, almost always. "Almost always" is used in the same sense as elsewhere in mathematics: there are at most countably many exceptions to an uncountable number of truths.

This is qualitatively and philosophically different from the converse idea that correlation implies causation, which is a logical error. Even if we assume that the presence of correlation between two variables implies the presence of causation, the correlation says nothing about the direction of causation, and furthermore, correlation can be explained by both variables being caused by some third variable. The next problem is measurement [10].

In the case of discrete variables, we run into a host of epistemological problems. The very first is that of natural kinds. If we want to measure the proportion of birds that fly, we need a criterion to distinguish birds from non-birds. As well, we need a definition of what it means to be able to fly. Even if this is possible, there is the problem of knowing when we have found all the entities we wish to call birds. There is also the fact that the world does not stay still while we are counting: as Bacchus [1] observes, we might do our counting in the spring, when most birds are flightless nestlings.

These problems get even more complicated once we try to measure continuous variables, whether with sticks and springs or laser beams. A practical theory of causation must address these questions of measurement, even if the ideas are theoretically robust.

However, to get any data in the first place (apart from ideally generated distributions), we must accept statistical, or *measurement*, correlation as a proxy for actual correlation. This may seem like a sleight-of-hand, but other methods for reliable causal inferences are subject to practical problems. For instance, Robins and Wasserman [19], state that

> We are not claiming that inferring causal relationships empirically is impossible. Randomized studies with complete compliance are a well-known example where reliable causal inference is possible.

The many practical difficulties inherent in collecting samples for studies, and many practical difficulties regarding compliance, in the view of Robins and Wasserman, do not undermine the idea of randomized studies enough to abandon the idea. (In fact, much simpler statistical inferences are subject to practical problems.)

A counterargument is that one may inspect samples and discard them if they are found not to be representative, and one may observe non-compliance. However, suppose the measured correlations are close to correct, but exist for the kind of reasons (i.e., cancellation) Cartwright posits. If that is the case, we still have a means for testing the predictive power of the theory. This is illustrated in the next section with an example using a causal visualization tool developed in our group.

7. Testing causal theories

A useful feature of causal theories is that they give us, under certain assumptions, the computational power to distinguish between *seeing* and *setting*. (Freedman [7] uses the terms *observation* and *intervention*.) *Seeing* is the common procedure of computing a conditional expectation – given a subpopulation of a known population, what is the posterior distribution of all related variables in the subpopulation? For example, we may wish to compute $p(M|\text{see}(S))$, the probability that someone we see using sunscreen might develop melanoma. This can be computed as the ordinary conditional probability $p(M|S)$.

Setting is about the consequences of actions (or interventions) and requires a subtler calculation. An individual wants to know the net effect of using sunscreen, for that individual, whether sunscreen use decreases or increases the overall probability of melanoma, in light of conflicting opinions. These two are computed as follows [16]:

$$p(M|\text{see}(S)) = p(M|DS)p(D|S) + p(M|\neg DS)p(\neg D|S),$$
$$p(M|\text{set}(S)) = p(M|DS)p(D) + p(M|\neg DS)p(\neg D).$$

Roughly, to compute the effect of setting S, we assume that the effect of sunscreen among all the Ds will have the same effect as it does among those who currently use sunscreen. (This explains the first multiplicative term in the second equation.) We make the same assumption for the $\neg D$s. Although this formulation goes back to [24], it has a simple implementation in causal graphs: erase arcs incoming to the *set* variable, and change all distributions involving the *set* variable accordingly. For correctness, see Pearl [16].

Suppose the resulting model is incorrect as a consequence of the case that causal relationships cancel each other out, showing *melanoma* and *sunscreen* as independent common causes of *darkness* (pigmentation) of skin, and mathematically is a minimal representation of possible causal relationships in a world coherent with the raw data. The user intuits this is incorrect and eventually constructs the correct model of causal relationships.

Fig. 4 shows the results of a user exploring *seeing* using a visualization tool we have developed [12]. The user grabs the value of the *sunscreen* variable, drags it up and down, and finds that *sunscreen* and *melanoma* are unrelated, as found by the original data mining tool (e.g., TETRAD). The statistical explanation is that *darkness* acts as a suppressor variable or confound. Light-skinned individuals are more likely to use sunscreen than dark-skinned people. However, they are also more likely to develop melanoma, exactly canceling the effect of the sunscreen. In this example, the distribution violates faithfulness

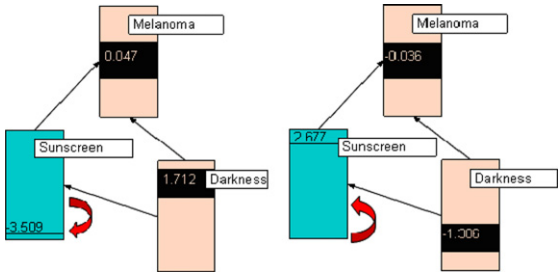

Fig. 4. Melanoma does not respond much to changes in sunscreen. Added exterior arrows indicate mouse movement during interaction.

– there is a spurious independency. (A recent study [5] found sunscreen positively associated with melanoma. We have changed the scenario for the present purpose.)

An experienced data analyst would revisit the data and ask what happens when *sunscreen* is manipulated after *darkness* is first fixed at some value Fig. 5 illustrates what needs to be done. The user first chooses to *see* a fixed value for the *darkness* variable, and then *sees* a range of values for *sunscreen by* dragging it up and down. The pairs of before and after images in Fig. 5 now reveal the correct relationship between *sunscreen* and *melanoma*. Whether *darkness* is high or low, fixing *darkness,* and then increasing *sunscreen* results in a decrease to *melanoma.*

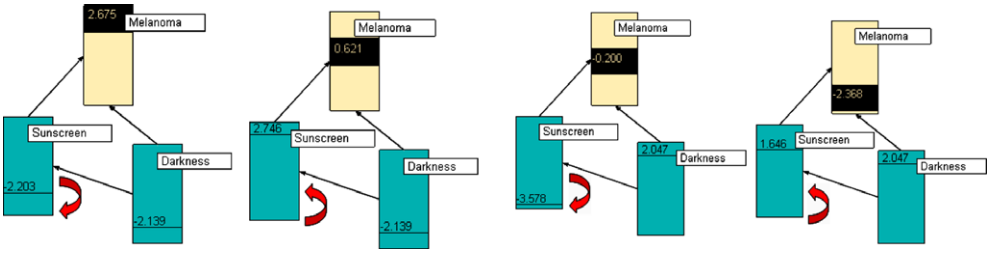

Fig. 5. Melanoma responds to seeing changes in sunscreen within a skin type. The top pair of graphs shows the change in melanoma with sunscreen for light-skinned persons, and the bottom pair shows melanoma change for dark-skinned persons. Ranges differ, but melanoma incidence consistently decreases with increased sunscreen usage.

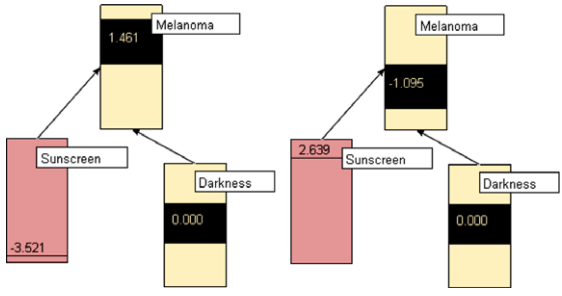

Fig. 6. Melanoma responds appropriately to setting of the sunscreen variable.

The user needs to perform a double *seeing* operation because there are two paths of probabilistic influence from *sunscreen* to *melanoma* that cancel each other out.

Because there are few potential confounds in this three node world, trying all *see* operations is not logistically difficult. However, in a richer dataset, this process may be cumbersome. *Setting* summarizes this combination of actions, as shown in Fig. 6.

Moreover, the predictions from the hypothesized cause-effect relationships give us something we can plausibly check in the real world.

8. Conclusions and ongoing work

We argue for carefully drawing a line between problems related to the theory of causal graphs and problems encountered with real world data. To paraphrase Druzdzel and Simon [6], the utility of any model is a function of the current state of knowledge, our ability to make accurate measurements, and the approximations we are willing to make.

Cartwright's factory example appears to be critical of causal reasoning formalisms because, to be effective, they require the imposition of the Markov condition prior to the determination of actual conditional independence relations and their depiction in a causal graph. We raised three questions in response to this. The first question was minor, regarding the distribution as presented by Cartwright. Our second question showed that the presence or absence of the Markov condition is in theory arbitrary. Lack of clear intuitions about its presence or absence can easily arise as a consequence of different measurement criteria. Finally, we believe that independence is not a consequence of the depiction, but rather something inherent in the depicted relations. The IC algorithm clearly depends on actual rather than assumed independencies.

We have replied to a criticism of faithfulness by suggesting that the intuitive semantic content of this assumption is that causation implies correlation. This idealization provides a basis for constructing and testing causal theories. We are exploring other characterizations. We have shown how predictions of such a theory can be explored, using a visualization tool that acts as a cognitive prosthetic [13]. These predictions can be tested either against our intuitions or with experiments or further measurements, and the results can be used to further constrain the data mining tool. By analogy, the IC algorithm relies on conditional independence information. Measure theoretically speaking, pure independencies are just as unlikely as the spurious independencies that violate faithfulness. More recent structure-from-data algorithms do not rely on categorical information about the presence of conditional independence and use scoring techniques to compute probabilities of different causal models. Similarly, Shipley [21] suggests statistical techniques for verifying causal models that contain such setpoints. In ongoing work, we are looking both at improving the visualization tool, and at other characterizations and criticisms of faithfulness.

Acknowledgements

This research was supported by a grant from the Natural Science and Engineering Research Council of Canada. Comments of some observant reviewers improved the presentation.

References

[1] F. Bacchus, A modest, but semantically well founded, inheritance reasoner, in: Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, 1989, Detroit, pp. 1104–1109.
[2] N. Cartwright, What is wrong with Bayes nets? in: H.E. Kyburg, Jr., M. Thalos (Eds.), Probability is the Very Guide of Life, 2003, OpenCourt Chicago, pp. 253–275.
[3] N. Cartwright, The Dappled World: A Study of the Boundaries of Science, Cambridge University Press, Cambridge, 1999.
[4] N. Cartwright, Causal diversity and the Markov condition, Synthese 121 (1999) 3–27.
[5] L.K. Dennis, L.E. Beane Freeman, M.J. VanBeek, Sunscreen use and the risk for melanoma: a quantitative review, Annals of Internal Medicine 139 (12) (2003) 966–978.
[6] M. Druzdzel, H. Simon, Causality in Bayesian belief networks, in: D. Heckerman, E.H. Mamdani (Eds.), Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence, 1993, Washington, pp. 3–11.
[7] D.A. Freedman, From association to causation via regression, in: V.R. McKim, S.P. Turner (Eds.), Causality in Crisis: Statistical Methods in the Search for Causal Knowledge in the Social Sciences, University of Notre Dame Press, South Bend, 1997, pp. 13–161.
[8] C. Glymour, Rabbit hunting, Synthese 121 (1999) 55–78.
[9] L. Kanal, L.J. Lemmer (Eds.), Uncertainty in Artificial Intelligence, North Holland, Amsterdam, 1986.
[10] H.E. Kyburg Jr., Theory and Measurement, Cambridge University Press, Cambridge, 1984.
[11] J. Lemmer, The causal Markov condition, fact or artifact? SIGART Bulletin 7 (3) (1996) 3–16.
[12] E. Neufeld, S. Kristtorn, Q. Guan, M. Sanscartier, C. Ware, Exploring causal influences, in: R.F. Erbacher, J.C. Roberts (Eds.), Proceedings of Visualization and Data Analysis 2005m, 2005, San Jose, pp. 52–62.
[13] E. Neufeld, D. Callele, D. Mould, S. Kristtorn, R. Rabu, A contribution to the theory and practice of cognitive prostheses, in: A.Butz, A. Krueger, P. Olivier (Eds.), Proceedings of Smart Graphics, 2003, pp. 241–250.
[14] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann Publishers Inc, San Francisco, 1988.
[15] J. Pearl, T. Verma, A theory of inferred causation, in: J. Allen, R. Fikes, E. Sandewall (Eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference, San Mateo, 1991 pp. 441–452.
[16] J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, New York, 2000.
[17] Y. Peng, J.A. Reggia, Plausibility of diagnostic hypotheses, in: T. Kehler, S. Rosenschein (Eds.) Proceedings of the 5th International Conference on AI, Philadelphia, 1986, pp. 140–145.
[18] H.E. Pople Jr., Heuristic methods for imposing structure on Ill-structured problems: the structuring of medical diagnostics, in: P. Szolovits (Ed.), Artificial Intelligence and Medicine, Westview Press, Boulder, Colorado, 1982, pp. 119–190.
[19] J. Robins, L. Wasserman, On the impossibility of inferring causation without background knowledge, in: C. Glymour, G. Cooper (Eds.), Computation Causation and Discovery, AAAI Press, California, 1999, pp. 305–321.
[20] R. Rodgers, C. Maranto, Causal models of publishing productivity in psychology, Journal of Applied Psychology 66 (1989) 688–701.
[21] B. Shipley, Cause and Correlation in Biology: A User's Guide to Path Analysis Structural Equations and Causal Inference, Cambridge University Press, 2000.
[22] P. Spirtes, C. Glymour, R. Scheines, An algorithm for fast recovery of sparse causal graphs, Social Science Computer Review 9 (1991) 62–72.
[23] P. Spirtes, C. Glymour, R. Scheines, Causation prediction and searchLecture Notes in Statistics, vol. 81, Springer-Verlag, New York, 1993.
[24] R.H. Stotz, H.O.A. Wold, Recursive versus nonrecursive systems: An attempt at synthesis, Econometrica 28 (1960) 417–427.
[25] T. Verma, J. Pearl, An algorithm for deciding if a set of observed independencies has a causal explanation, in: D. Dubois, M.P. Wellman (Eds.), Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence, 1992, pp. 323–330.