An Information-theoretic Approach to Causal Power

Technical Report 2005/176

Lucas R. Hope and Kevin B. Korb
Clayton School of Information Technology
Monash University
Clayton, Victoria
Australia.

Abstract

The measurement of causal power has been of long-standing interest in philosophy and psychology. With the development of Bayesian network technology, the interest has spread to artificial intelligence. We introduce a metric which applies information theory to Bayesian networks. We show that it generalizes prior metrics restricted to linear and noisy-or models, while providing a metric appropriate to the full representational power of Bayesian nets.

Keywords: Bayesian networks, ontology, information theory, knowledge engineering.

1 Introduction

The causal interpretation of Bayesian networks has risen greatly in prominence since the development of effective probabilistic reasoning with Bayesian networks (Pearl, 1988), and more especially since the development of methods of learning Bayesian networks from data in causal discovery algorithms (Verma and Pearl, 1990; Spirtes et al., 2000; Neapolitan, 2004). The causal interpretation makes sense of the success of causal discovery, which otherwise would be miraculous. Despite this, the causal interpretation brings with it a host of difficulties, philosophical and technical, leading to various research efforts, such as the attempt to apply the theory of probabilistic causality in philosophy of science to causal Bayesian networks (e.g., Hitchcock, 2001; Halpern and Pearl, 2001; Twardy and Korb, 2004). Another long-standing research program in philosophy and psychology has attempted to develop a formal theory of causal power: as probabilistic causality already recognizes, causation is not an either/or affair, but comes in degrees. But then explanatory power — the normative attribution of an effect to one of its causes — ought also to come in degrees, particularly in cases of overdetermination or multiple causation. The development of a well-founded metric of causal power promises to be of wide-spread interest: as a normative standard for assessing such causal attributions; as an aid in knowledge engineering Bayesian networks, by providing guidance in the interpretation of prototype networks; as an aid in understanding and using probabilistic expert systems, in simplifying and making explicit causal relationships that are otherwise complicated or implicit; and also for the growing collaboration between AI and philosophy of science, in understanding, for example, the nature of scientific explanation.

Here we review the best known prior theories of causal power, beginning with I.J. Good (1961) and then the recent work of Cheng (1997), Glymour (2001) and Hiddleston (2005). The main problem we find in all of these theories is that they find their inspiration in an overly simple model of causality, namely linear (or additive) causal models. Linear causality is a particularly simple kind of causality, which ought to have a correspondingly simple account to offer of causal power. Whereas simplicity can be an asset in first developing a theory, it can be a serious impediment when attempting to generalize a theory, and we suggest that that is the predicament of causal power theory. In particular, the transitive nature of causality in linear models has seduced some into thinking that causality is in general transitive. However, it is not, as Christopher Hitchcock (2001) and others have shown. In response to this we offer an account of causal power using an information-theoretic analysis of the relationships in non-linear Bayesian networks, while also illustrating their application to linear models.

2 Desiderata of a Causal Power Theory

There are a number of principles we would like a causal power theory to exemplify.

1. The theory should be applicable to linear path models. This means we'll have to abandon — or reinterpret substantially — theories which have probability contrast at their core. Furthermore, the theory should bear a direct relationship to linear correlation.

2. The theory should have an information-theoretic interpretation. Causality gives rise to probabilistic relationships, which should lead to a reasonable interpretation under Shannon's theory of information.

3. The theory should not depend on transitivity, simply because causation is not, in general, transitive; of course, the theory needs to accommodate transitivity when it appears.

4. The theory should be compatible with experimental intervention. Thus, we should be able to quantify the effect of setting a variable to a particular value, as well as contrasting the effect of one value with another.

In describing the various theories of causal power we shall adopt the standard convention of writing variables as capital letters (e.g., C for a causal variable and E for an effect variable) and the states or values they take in lower case. Thus, $C = c$ means we are considering the variable C in its particular state c. And c alone may be used as shorthand for $C = c$, etc.

3 A History of Causal Power

3.1 Good's Causal Calculus

The first serious attempt to provide a formal theory of causal power that we know of is that of I.J. Good (1961). In the same work, Good also presented the first probabilistic theory of causality (see Salmon, 1980, for a review).

Good's formulation of causal power seems motivated by a desire for a theory analagous to circuit theory. Causal strength (Q) is analogous to conductivity, and he defines a kind of 'causal resistance' (R) to parallel circuit resistance. In circuits, resistors in series are additive; in turn, Good's causal resistance is additive along a causal chain. Conductivity and causal power, on the other hand, are additive in parallel. In circuits, conductivity is the reciprocal of resistance; similar to this, Good's causal strength and causal resistance are related thus:

$$e^{-R} + e^{-Q} = 1 \tag{1}$$

Given these requirements, the only additional requirement is a definition either of causal strength or of causal resistance for direct causal links (which we will call Q_{link}). Good's definition of causal strength for a causal link is:[1]

$$Q_{link}(E:C) = -\log \frac{P(\neg e|c)}{P(\neg e|\neg c)} \tag{2}$$

$$= -\log \frac{1-p}{1-q} \tag{3}$$

where $p = P(e|c)$ and $q = P(e|\neg c)$. Good stipulates that Q_{link} be non-negative: where the formula above would yield a negative value, it takes zero instead. Thus, c must promote e for Q_{link} to be non-zero. Good calls this formula "the weight of evidence against e, if c did not happen."

Using the fact that resistance is additive along a chain, together with formula (1) relating resistance and causal strength, it follows that:

$$Q(E:C) = -\log \left(1 - \prod_i \frac{p_i - q_i}{1 - q_i}\right) \tag{4}$$

where c and e are connected by a chain of links indexed by i. This formula is derived by finding each Q_{link}'s corresponding R_{link} using (1), summing the R's to get the total resistance, then reapplying (1) to transform the sum into a causal power expression.

Good's theory has some nice properties; the analogy to circuit resistances in particular is mathematically pleasing, as is the use of information-theoretic ideas. However, there are some key objections. The first is that the theory exhibits a form of transitivity, because of the additivity of resistances. Since causation in general is not transitive, this will often yield the wrong answer. Take, for example, Richard Neapolitan's case of finesteride (Neapolitan, 2003). Finesteride reduces testosterone levels; lowered testosterone levels can lead to erectile dysfunction. However, finesteride fails to reduce testosterone levels *sufficiently* for the follow-on erectile dysfunction to occur.[2] Good's calculus of causal strength cannot represent these facts: as soon as the individual Q_i in a chain all are given positive values, the end-to-end Q, by formula (4), also must take a positive value. Salmon (1980) also pointed out technical difficulties in Good's calculus which allow distinct causal chains with distinct end-to-end dependencies to be accorded the same end-to-end Q values, evidently misrepresenting the causal story.

[1] Good includes the context in his formula, which we leave implicit here.

[2] This was reported in at least one study. Whether or not true generally, the point is that it *could* be true. Incidentally, this case does *not* appear in Neapolitan (2004) because the publisher thought the example too challenging for its delicate readership!

3.2 Cheng's Power PC Theory

The starting point for the probabilistic theory of casuality is that a positive probabilistic dependency is a necessary condition of causality: no causation without probabilification. This can be represented by finding what Cheng calls a positive (or negative, given the opposite sign) probabilistic contrast: $\Delta P_c = P(e|c) - P(e|\neg c) > 0$ — which Suppes (1970) called *prima facie causation*. In this case c is only a *prima facie* cause, since a common ancestor may be responsible for correlating two effects; so Suppes' theory went on to lay down conditions ruling such cases out. The later research program in probabilistic causality is largely concerned with further refinement of such conditions, which in the end have been subsumed by technical developments in the Bayesian network theory for representing conditional independence (cf. Twardy and Korb, 2004).

Cheng's causal power theory also attempts to overcome the limitations of *prima facie* causation, which she calls candidate generative (or preventive) causation.

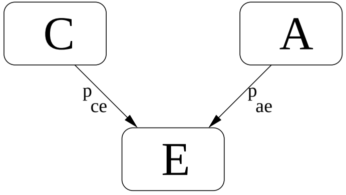

Figure 1: The structural restrictions implied by Cheng's power PC.

3.2.1 Generative causes.

Cheng's causal power theory begins with some very stringent requirements for causal structure. The covariation between the effect e and candidate cause c must be independent from any covariation of e and all other causes (gathered together as a, in Figure 1). Further, the occurence of c must itself be independent of a. This implies either the limited causal structure represented by Figure 1 or else that all the causes of a are fixed.

Cheng defines the causal power of c for e as the probability that c produces (or generates) e. The causal power of c is labelled p_c, leaving e implicit. Under Cheng's assumptions, e can only be caused by either a or c, so $P(e)$ can be written as:

$$P(e) = P(c)p_c + P(a)p_a - P(c)P(a)p_c p_a \tag{5}$$

Conditioning on c and $\neg c$,

$$P(e|c) = p_c + P(a|c)p_a - p_c P(a|c)p_a \qquad \text{(as } P(c|c) = 1\text{)} \tag{6}$$

$$P(e|\neg c) = P(a|\neg c)p_a \qquad \text{(as } P(c|\neg c) = 0\text{)} \tag{7}$$

This allows her to redefine ΔP using p_c and p_a:

$$\Delta P_c = P(e|c) - P(e|\neg c) \tag{8}$$

$$= p_c + P(a|c)p_a - p_c P(a|c)p_a - P(a|\neg c)p_a \tag{9}$$

$$= [1 - P(a|c)p_a]p_c + [P(a|c) - P(a|\neg c)]p_a \tag{10}$$

Finally, this can be solved for p_c:

$$p_c = \frac{\Delta P_c - [P(a|c) - P(a|\neg c)]p_a}{1 - P(a|c)p_a} \tag{11}$$

This equation requires p_a, so we use the above assumptions to eliminate it, as follows. Assuming a occurs independently from c means that $P(a|c) = P(a|\neg c) = P(a)$, and so the p_a term in the numerator of (11) disappears. This, together with the assumption that a's effect on e is independent from c, lets us use $P(e|\neg c)$ to estimate $P(a|c)p_a = P(a)p_a$. This is because in the cases where c is not present, a is the only generative cause of e by definition.

Thus, under the above assumptions we have:

$$p_c = \frac{\Delta P_c}{1 - P(e|\neg c)} \tag{12}$$

Cheng claims that this is a significant improvement on previous theories, such as that of Rescorla and Wagner (1972) because, among other reasons, it provides the correct answer for when e always occurs. If e always occurs, then the value for p_c is undefined, rather than the zero Rascorla and Wagner suggest. Leaving p_c unspecified is deemed correct because we should be unable to assess the causes of a universal event.

3.2.2 Preventive causes.

Cheng stipulates the same structural restrictions for preventive causes as she does for generative ones. Her definitions are the same as for generative causes, except now p_c is preventive, thus p_a is the only generative cause; to distinguish prevention notationally, we will write such powers as $\bar{p}_c$.

Cheng takes the occurrence of e to be the combination of e produced by a with e *not being stopped* by c, and therefore:

$$P(e) = P(a)p_a(1 - P(c)\bar{p}_c) \tag{13}$$

This means that e being produced by a occurs independently from e being prevented by c, which is a rather strange idea, as noted by Hiddleston (2005), and discussed further in Section 3.4.

As with generative causes, this is used to find $P(e|c)$, $P(e|\neg c)$ and ΔP_c:

$$P(e|c) = P(a|c)p_a(1 - \bar{p}_c) \tag{14}$$

$$P(e|\neg c) = P(a|\neg c)p_a \tag{15}$$

$$\Delta P_c = P(a|c)p_a - P(a|c)p_a\bar{p}_c - P(a|\neg c)p_a \tag{16}$$

(16) is solved for $\bar{p}_c$ to obtain:

$$\bar{p}_c = \frac{[P(a|c) - P(a|\neg c)]p_a - \Delta P_c}{P(a|c)p_a} \tag{17}$$

Cheng then uses the same assumptions as for generative causes to get:

$$\bar{p}_c = \frac{-\Delta P_c}{P(e|\neg c)} \tag{18}$$

Analogously to positive causation, this leaves the preventive power of a for a never-occurring e undefined.

3.2.3 Problems with Cheng's theory.

The main difficulty for Cheng's theory is that it is extremely limited in scope. In addition to being defined only over binomial variables, it has severe limitations in scope, due to both the structural restrictions and the restrictions on interactions between variables. These are necessary to make her derivations work, but constrict the theory to a very limited range of cases.

Novick and Cheng (2004) extend the model to deal with a limited form of interaction between causes. Briefly, the causal power of an interaction between causes c_1 and c_2 is calculated by finding the difference between the actual probability $P(e|c_1, c_2)$ and the counterfactual probability $P'(e|c_1, c_2)$ which assumes there is no interaction between c_1 and c_2. This extension does not relax the structural requirements of the original formulation, so we do not deal with it further here.

3.3 Glymour's Extension to Cheng Models

As noted in Section 3.2.3, Cheng's (1997) power PC models require that candidate causes occur independently of each other. Glymour (2001) extends Cheng's models to a limited set of Bayesian models. His extension assumes non-interacting causes, so the work of Novick and Cheng (2004) does not apply.

Glymour has derived a set of structural rules that are more general than Cheng's, under which the power PC theory holds. We paraphrase them here:

1. If an effect e has a single generating cause a and an independent unobserved preventer u, then the causal power of a *cannot* be estimated.

2. If e has any number of observed generating and preventing causes, and an independent unobserved preventing cause, then the ratios of any two generating causes can be estimated.

3. If e has at least one observed generating cause and an independent unobserved preventing cause, observed preventing causes can be estimated.

4. If e has an independent generating cause, then the causal power of an observed cause (whether generating or preventing) can be estimated.

5. If e has an unobserved generating cause u, a generating cause a may be estimated if (1) u does not cause a and (2) another observed cause b of e is not directly caused by u *as well as* either caused by a or connected via an unobserved common cause to a.

The first rule shows a particular case where the causal power cannot be determined, while the rest detail cases where it can be found. The second through fourth rules describe situations where the unobserved cause is *independent* from the other causes. The last rule describes a complicated case where the unobserved cause is confounding other observed causes. For the particular confounding case presented, causal power can be estimated.

Glymour has extended Cheng models in two ways. First, he notes that her equations are easily adapted to Bayesian Networks, albeit ones with constrained probability relations (that is, causes must not interact). He also detailed a set of rules for when the causal power can be assessed when an unobserved cause is present.

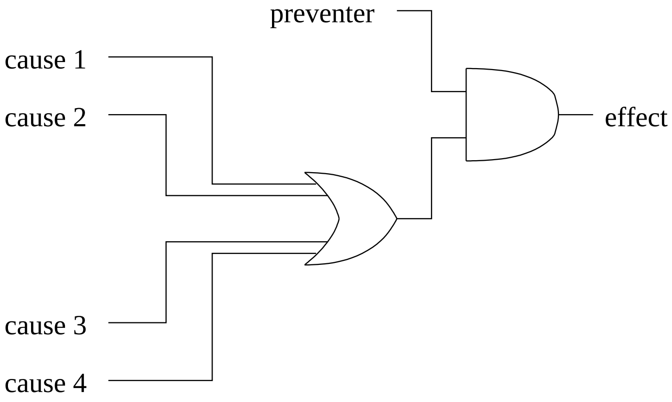

Figure 2: How Cheng's preventive powers work.

3.4 Hiddleston's Causal Powers

Hiddleston's (2005) analysis of causal powers is heavily influenced by Cheng's and Glymour's accounts. However, he disagrees with Cheng's formulation of preventive causes. Recall Cheng's formula (13) for how e occurs when c is a preventive:

$$P(e) = P(a)p_a(1 - P(c)\bar{p_c}) \tag{19}$$

This means that e occurs only when a causes it and, independently, c fails to prevent it. But Hiddleston argues that preventers work by preventing particular causes, and so he suggests replacing Cheng's formula with:

$$P(e) = P(a)p_a(1 - P(c|a)\bar{p_{c,a}}) \tag{20}$$

where $\bar{p_{c,a}}$ is c's probability of preventing a's effect on e.

This difference between Cheng's and Hiddleston's accounts can be thought of as a difference between two kinds of barrier against generative powers. Cheng's is a uniform barrier against all possible generative causes (as in Figure 2), while Hiddleston's only shields against a specific cause.

4 Wright's Method of Path Coefficients

Wright (1934) was the first to use diagrams for representing causal relationships between random variables, although limiting them to linear causal relationships. Wright began with a formula relating deviations from the mean and correlation:

$$(V_0 - \bar{V_0}) = c_{01}(V_1 - \bar{V_1}) + c_{02}(V_2 - \bar{V_2}) + \ldots + c_{0n}(V_n - \bar{V_n}) \tag{21}$$

Where $V_0, \ldots, V_n$ are variables, $\bar{V_i}$ is the mean of V_i, and $c_{01}, \ldots, c_{0n}$ are the correlation coefficients between V_0 and the other variables. If we set $X_i = (V_i - \bar{V_i})/\sigma_i$, the standardized deviation of V_i, then the path coefficient is $P_{0i} = c_{0i}(\sigma_i/\sigma_0)$, which measures the (normalized) proportion of the deviation of V_0 contributed by V_i. The above equation can then be rewritten as:

$$X_0 = P_{01}X_1 + P_{02}X_2 + \ldots + P_{0n}X_n \tag{22}$$

Wright's path diagrams contain directed arcs between variables, representing direct causal relationships, and undirected (or bi-directed) arcs, representing correlations not further analysed. The correlation between any two variables is the sum of the correlations attributable to each path between them, with two sorts of paths being inadmissable. If two variables directly affect another variable, then no correlation between the parents can be ascertained from this connection (you cannot trace forward and then backwards through the diagram). Similarly, only one undirected (residual) correlation can be included in a path.

The correlation r_{0g} between variables X_0 and X_g can be calculated recursively:

$$r_{0g} = \sum_i P_{0i} r_{gi} \tag{23}$$

Note that if X_0 and X_g are directly connected (i.e., one is a descendant of the other), then X_0 must be the descendant.

Wright does not discuss causal power as such, but a causal power metric is easily developed for his theory. We start, as does Wright, with all arcs oriented in the causal direction, and replace Wright's restrictions with a stronger one: paths containing either common causes or common effects are inadmissable (so you cannot trace forward then backwards *or* backwards then forward). This can be implemented by simply cutting the arcs between the cause and its parents, as in current interventionist accounts of causality, which model causal interventions by cutting the arcs between the intervened-upon variable and its parents (e.g., Pearl, 2000a). The correlation that remains, as given by Wright's equation, measures the causal power of the cause under test and the effect, across all directed paths from the former to the latter.

Unfortunately, undirected paths are a problem. In our account here, the causal power associated with these paths is undefined. And if we cannot assess the causal power of a single path, then the total causal power cannot be assessed. Our solution is simply to disallow undirected paths in causal path models. This is not merely because our account cannot accommodate them: directed paths are metaphysically fundamental; correlations associated with undirected paths are just those whose underlying causal story has yet to be told, whether by orienting arcs or by the discovery of latent variables.

Wright's method of path coefficients is simple and elegant, and can be readily extended to incorporate a theory of causal power. The only disadvantage is its limitation to linear relations.

In any case, this theory provides a litmus test for more general theories. If a general theory, in its restriction to path models, does not provide the same answers as Wright's method, then we shall hold it to be in error.

5 Causal Information and Causal Power

We now develop a theory of causal power which blends information theory and causal intervention theory (Pearl, 2000b; Korb et al., 2004).

The basic case, which the theories of causal power described above treat, is the power of some specific event to cause some other specific event; it thus concerns particular values of the cause and effect variables. Information theory, however, relates pairs of variables to each other, across the full range of values they may take, rather than for any particular pair of values of those variables. Therefore, we shall first introduce an information-theoretic metric that relates variables, which we call **causal information** and then particularize the metric to relate values of variables, rendering the concept relevant to the kinds of cases discussed above, and which we call **causal power**.

5.1 Causal Information

Definition 1 *The* **causal information** *(CI) between a cause C and an effect E in the causal model g is*

$$CI(C,E) = k \sum_{c \in C, e \in E} p(e|c) \log \frac{p(e|c)}{p(e)} \tag{24}$$

with some parameter $k \in R^{+}$. This is to be measured between the two variables in the auxiliary model g^{*}, where g^{*} is the same as g, except the arcs between C and its parents have been cut (removed).

This is closely related to *mutual information*, which has the definition for the discrete case (Cover and Thomas, 1991):

$$MI(X,Y) = \sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x|y)}{p(x)} \tag{25}$$

When parameter k is set to $1/|C|$, causal information is equal to mutual information (in the auxiliary model) when $p(C)$ is uniform. This setting is convenient and we use it throughout this paper.[3]

We can specialize the definition of causal information along either, or both, of two dimensions, namely relative to a particular causal event or relative to a particular effect event.

Definition 2 *The* **causal information** *(CI) between a cause $C = c$ and an effect E in the causal model g is*

$$CI(C=c,E) = k \sum_{e \in E} p(e|c) \log \frac{p(e|c)}{p(e)} \tag{26}$$

with some parameter $k \in R^{+}$, as measured in the auxiliary model g^{*}.

This causal information describes the power of a particular value c of C to influence the effect variable E in general. We may equally well be interested in the influence of C in general upon a particular state of E.

Definition 3 *The* **causal information** *(CI) between a cause C and an effect $E = e$ in the causal model g is*

$$CI(C,E=e) = k \sum_{c \in C} p(e|c) \log \frac{p(e|c)}{p(e)} \tag{27}$$

with some parameter $k \in R^{+}$, as measured in the auxiliary model g^{*}.

[3] If k is set instead to the actual prior probability $p(c)$, then this becomes causal information for interventions, as we discuss below.

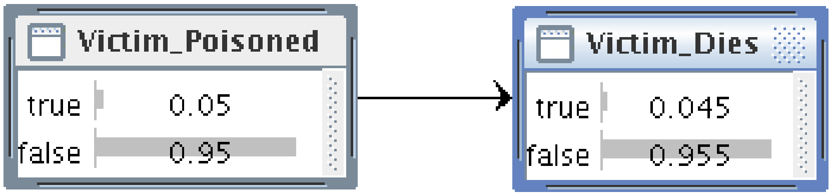

Figure 3: The poison example. Mutual information changes greatly depending on the probability of *Victim Poisoned*, whereas causal information is invariant.

5.2 Causal Power

Finally, we may be interested in the very specific question of how $C = c$ is related to $E = e$. This is the most typical question raised, in fact, in circumstances of causal explanation: when two particular events have occurred, how much "responsibility" for the effect's value can we attribute to the cause's value? We use the designation "causal power" for these cases, using "causal information" as the more general term.

Definition 4 *The* **causal power** *(CP) between a cause $C = c$ and an effect $E = e$ in the causal model g is*

$$CP(c, e) = kp(e|c) \log \frac{p(e|c)}{p(e)} \tag{28}$$

with some parameter $k \in R^+$, as measured in the auxiliary model g^*.

This account of causal power has the advantage of providing a unified treatment of both generating and preventive causes. The same formula applies in either case, with CP going negative when the particular causal state prevents the effect.

5.3 Causal Power and Mutual Information

One might wonder why we do not just use mutual information as our measure of causal relationship. Although mutual information is symmetric, its use for causal information cannot be, since the measurements of Definition 1 are required to occur in the (asymmetric) auxiliary model, rather than in the original model. But another difficulty is that mutual information depends upon the prior probability of the cause. Causal power (and information) is surely relative to *background* contexts: for example, the relevance of an antidote to health is very low generally, but high when a particular toxin is present. Thus, we expect the probability distributions to be used in applying any metric shall be relativized to some set of background observations. However, it is not so sensible to relativize a causal power metric to a set of conditions yielding a specific probability distribution over the cause itself. If a highly effective poison is applied to a victim, that will be a powerful explanation of the victim's subsequent death, particularly if no other explanations are plausible in the circumstances. But then the powerful explanation will not at all be undermined should closer scrutiny reveal that the circumstances were highly likely to lead to just this kind of poisoning! In general, causal power should not depend upon the chance of the cause occurring, but rather upon its role in bringing about the effect *should* it occur.

In the example of Figure 3 poison has a 0.9 probability of causing death — i.e., $P(death|poison) = 0.9$. Its causal information is 0.758, whereas the mutual information varies from zero to half this causal information, depending upon the probability of the cause.

5.4 The Causal Power of Interventions

Despite these remarks, there is some role for the assessment of causal power relative to the probability of the cause itself, measured by Equation (28) with $k = p(c)$. We can call this particular CP the *causal power of an intervention* on C. For example, if a public health official is considering the merit of administering a vaccine to a group of people, rendering them immune to a threatened disease, it will be important to take into account not just the efficacy of the vaccine to induce immunity (as well as side-effects, etc.), but also the prior probability of individuals *already* being immune (having already received the vaccine). This is so precisely because in such a case it is not the isolated efficacy of the cause which is at issue, but the efficacy of the intervention upon the cause.

5.5 Information-theoretic Interpretations

As we have already noted, causal information can be viewed as a special kind of mutual information. It can also be interpreted in terms of Kullback-Leibler (KL) divergence (or cross-entropy). KL divergence takes the form

$$KL(p(X), q(X)) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \tag{29}$$

where p is taken to be the true distribution and q an approximation to p. KL is a measure of the expected information cost of using distribution q to describe p. Mutual information can be understood as the information cost of assuming the two variables X and Y are independent when they may not be; that is, it is equal to the KL divergence from $p(X,Y)$ to $p(X)p(Y)$.

The KL divergence interpretation of causal information, then, is just that the causal information of c for E is the KL divergence from $p(E|c)$ to $P(E)$.

Theorem 1 *The causal information of $c \in C$ for E is:*

$$CI(C = c, E) = k \sum_{e \in E} p(e|c) \log \frac{p(e|c)}{p(e)} = KL(p(E|c), p(E)) \tag{30}$$

with the last equality holding if we set $k = 1$.

This theorem follows immediately from Definition 2 and (29).

This account of causal power has the immediate advantage of being defined in general, applying to any probabilistic system for which we can find the underlying causal structure.[4] (The causal structure is necessary in order to identify which arcs are to be cut under intervention, of course.) Thus, it applies to linear models, Cheng models and their extensions, and additionally to discrete variable models representing interactive causes — applying to the full range of (causal) Bayesian networks, unlike any of its predecessors. In order to assess this account against its predecessors, however, we need to look at how it applies to the simpler cases of linear and Cheng models, which we now do.

[4] This includes being defined over continuous probability spaces, where probabilities, mutual information, entropy, etc. are redefined in terms of integrals, rather than sums. The definition of causal information is analogously redefined in such cases.

6 Applications

6.1 Path Models

To test the application of causal information to linear models we consider it in connection with Wright's theory of path models, which are a general method of treating linear Gaussian (normal) models and which apply the unit normal $N(0,1)$ to all residual terms by standardizing all variables. The particular requirement is that causal information make sense of correlation, as calculated by the method of Wright (1934), which we do by treating causal information in the form of mutual information.[5]

To find the mutual information between two correlated normals, it is easiest to start with the identity $MI(X,Y) = H(X) + H(Y) - H(X,Y)$, where H is the entropy. The unit normal distribution $N(0,1)$ has the density:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \tag{31}$$

The entropy for a continuous variable is:

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx \tag{32}$$

Therefore the entropy for $N(0,1)$ is:

$$\begin{aligned}
H(N(0,1)) &= -\int_{-\infty}^{\infty} f(x) \log\left[\frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\right] dx && (33)\\
&= -\int_{-\infty}^{\infty} f(x) \left[-\log\sqrt{2\pi} - \frac{x^2}{2}\right] dx && (34)\\
&= \frac{1}{2}\log 2\pi \int_{-\infty}^{\infty} f(x)dx + \frac{1}{2}\int_{-\infty}^{\infty} f(x)x^2 dx && (35)\\
&= \frac{1}{2}\log 2\pi + \frac{1}{2} && (36)\\
&= \frac{1}{2}\log 2\pi e && (37)
\end{aligned}$$

since $\int_{-\infty}^{\infty} f(x)dx = 1$ and $\int_{-\infty}^{\infty} f(x)x^2 dx = E(x^2) = \sigma^2 = 1$.

The joint entropy for two continuous variables is:

$$H(X,Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) \log f(x,y) dy dx \tag{38}$$

Two normal distributions have the following joint density, where $\mu_x = \mu_y = 0$ and $\sigma_x = \sigma_y = 1$ (as is assumed in path models), with r being the correlation.

$$f(x,y) = \frac{1}{2\pi\sqrt{1-r^2}} e^{-\frac{x^2+y^2-2rxy}{2(1-r^2)}} \tag{39}$$

[5] Wright details a recursive method of calculating total correlation from causal (directed) paths.

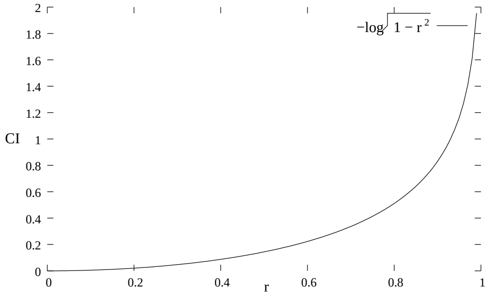

Figure 4: Mutual information is an increasing function of correlation.

From (38) the joint entropy for X and Y is:

$$(40) H(X,Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\log\left[\frac{1}{2\pi\sqrt{1-r^2}}e^{-\frac{x^2+y^2-2rxy}{2(1-r^2)}}\right]dydx$$

$$\begin{aligned}&= \log 2\pi\sqrt{1-r^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy\\ &\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{x^2+y^2-2rxy}{2(1-r^2)}f(x,y)dxdy\end{aligned}\tag{41}$$

$$\begin{aligned}&= \log 2\pi\sqrt{1-r^2} + \frac{1}{2(1-r^2)}\left[\int_{-\infty}^{\infty} x^2 f(x)dx + \int_{-\infty}^{\infty} y^2 f(y)dy\right.\\ &\qquad\left. - 2r\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy\right]\end{aligned}\tag{42}$$

$$= \log 2\pi\sqrt{1-r^2} + \frac{1+1-2r^2}{2(1-r^2)}\tag{43}$$

$$= \log 2\pi e\sqrt{1-r^2}\tag{44}$$

Since $E(X^2) = E(Y^2) = \sigma_x^2 = \sigma_y^2 = 1$, and $E(XY) = r$ by definition.

Now that we have the values of $H(X,Y)$, $H(X)$ and $H(Y)$, we can calculate $MI(X,Y)$:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y)\tag{45}$$

$$= \frac{1}{2}\log 2\pi e + \frac{1}{2}\log 2\pi e - \log 2\pi e\sqrt{1-r^2}\tag{46}$$

$$= -\log\sqrt{1-r^2}\tag{47}$$

The relationship between mutual information and correlation is shown in Figure 4. This relationship is pleasingly simple and in line with our intuitions, mutual information is a monotonically increasing function of correlation magnitude. As we should expect and demand for linear models, the causal information is transitive, since correlation is transitive.

6.2 Causal Networks

6.2.1 Bayesian networks.

Mutual information has already been used with Bayesian networks (BNs) in a number of ways, for example, in performing sensitivity analysis. BNs were originally designed to be concise representations of a joint probability distribution over a number of random variables. A Bayesian network represents the existence of a direct statistical dependency between each pair of random variables by a directed arc (or arrow). The resultant graph then defines a factorization of the joint distribution using conditional probability tables given for each variable in terms of its parents.

Under this interpretation of BNs, causality plays no part. Since our concerns are causal, we add the constraint that arcs are oriented causally: they go from switch to light bulb, and never the other way around. The resultant networks are then properly called causal networks. These networks are generally discrete in practice, because the most popular algorithms for probability updating need them to be so. Even given these constraints, causal networks are quite general constructs; they can represent any discrete probability distribution or function. So for instance, they subsume the models of Sections 3.2-3.4. Given the right relationship between the variables (i.e., the right conditional probability table), any of those models can be emulated by a causal network.

6.2.2 Applying causal information to Cheng models.

We described in sections 4 and 5 roughly how to calculate causal information. More precisely, the steps are:

1. Obtain the causal model for the domain.[6]

2. Select an observational context within which the causal question is to be asked and answered. E.g., the context in the antidote case may specify whether or not poison is present. What variables *cannot* be allowed into the context are either descendants of the effect variable (which improperly give evidence about that effect) or variables lying on a directed path between the cause and the effect (unless specifically the causal power restricted to the remaining paths is what is of interest).

3. Cut all arcs into the cause.

4. Compute the causal information.

Disregarding prevention *pro tem*, Cheng models are the pychological equivalent of the *noisy-or* relationship between binary variables. It is easiest to calculate noisy-or using the probabilities of the effect being inhibited, that is, probabilities of the form $q = 1 - p$, where p is the probability of the effect given some cause. If $pa(e)$ are the parents of e and $pa_T(e)$ the subset of parents which are true, then,

$$p(e|pa(e)) = 1 - \prod_{i \in pa_T(e)} q_i \tag{48}$$

where q_i are the probabilities that the individual causes fail to bring about the effect (i.e., the probability that the i-th "inhibitor" is active). The probability of e being false is the probability that all the inhibitors of the occurrent causes activate. Since the inhibitors are assumed to be independent, this is the product of their individual probabilities, so the probability of e is just one minus this quantity.[7]

Prevention can also be modelled by a sort of *noisy-and* relationship, although this is not as nice a fit as noisy-or for generation. We use Hiddleston's prevention (see Section 3.4)

[6] Questions about how to perform this complicated step are outside the scope of this paper; see, e.g., Verma and Pearl (1990); Spirtes et al. (2000); Neapolitan (2004); Korb and Nicholson (2004).

[7] Some specify a 'leak' probability in the node, which is the chance that the effect will occur due to unspecified causes. We ignore leak probability here, noting that it can be obtained by specifying an extra parent of e representing all other causes and forcing it to be true.

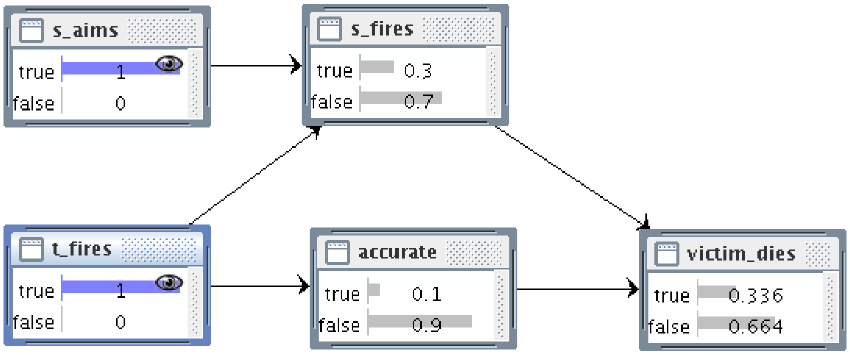

Figure 5: The sharpshooter example: Trainee fires is actually preventive in the case where sharpshooter aims, because it prevents the highly accurate sharpshooter from firing. (The eyes indicate observed variable states.)

because preventers targeting individual causes work better here. His theory is also more general, because it can represent Cheng's prevention by gathering individual causes together (through an or-gate) before applying prevention.

Recall equation (20):

$$P(e) = P(a)p_a(1 - P(i|a)p^-_{c,a})$$

where $p^-_{c,a}$ is c's probability of preventing a's effect on e.

We shall now compare Cheng's theory of causal power and our own in some examples.

6.2.3 Example: Sharpshooter.

Hiddleston's sharpshooter case contains a preventive cause as well as generative causes. The idea is that a sharpshooter is supervising a trainee assassin, having the job of shooting the victim if the trainee loses her nerve and fails to shoot. The sharpshooter either does or doesn't aim at the victim in advance and has a better chance of hitting the victim if he does, should he fire. If the trainee shoots at the victim, she has a good chance of preventing the sharpshooter from firing. What might be surprising, the trainee's taking a shot is overall good for the victim's health, i.e. the victim is more likely to die if the trainee does not shoot and so (very likely) the sharpshooter does. This ensues when the trainee is significantly more inaccurate than the sharpshooter.

Figure 5 shows the scenario where the trainee fires and the sharpshooter aims.[8] In this case, the ΔP of the trainee firing for the victim's death is $0.336 - 0.9 = -0.564$, so the preventive power $p^-_{c,a}$, taking sharpshooter aims as the generative cause, is

$$p^-_{c,a} = \frac{0.564}{0.9} = 0.627 \quad \text{(the sharpshooter aims)} \tag{49}$$

$$p_{c,a} = \frac{0.09}{0.99} = 0.091 \quad \text{(the sharpshooter does not aim)} \tag{50}$$

Note that the former employs (18), being a prevention case, while the latter employs (12).

[8] Note that the conditional probability tables are not shown in the figure, but only the event probabilities under the circumstances given.

The causal information results are:

$$\text{(51)} \qquad \mathrm{CP}(\mathrm{c},\mathrm{e}) = \frac{0.336}{2}\log\frac{0.336}{0.618} = -0.148 \quad \text{(the sharpshooter aims)}$$

$$\text{(52)} \qquad \mathrm{CP}(\mathrm{c},\mathrm{e}) = \frac{0.09}{2}\log\frac{0.09}{0.0495} = 0.0388 \quad \text{(the sharpshooter does not aim.)}$$

Both accounts agree on the most basic feature of the example: the trainee firing prevents death when the sharpshooter is aiming and otherwise causes death. The ratio between the two powers for Cheng is 6.89, while for our causal power it is −3.81. So both confirm that the trainee's effect on the victim is larger when the sharpshooter aims.

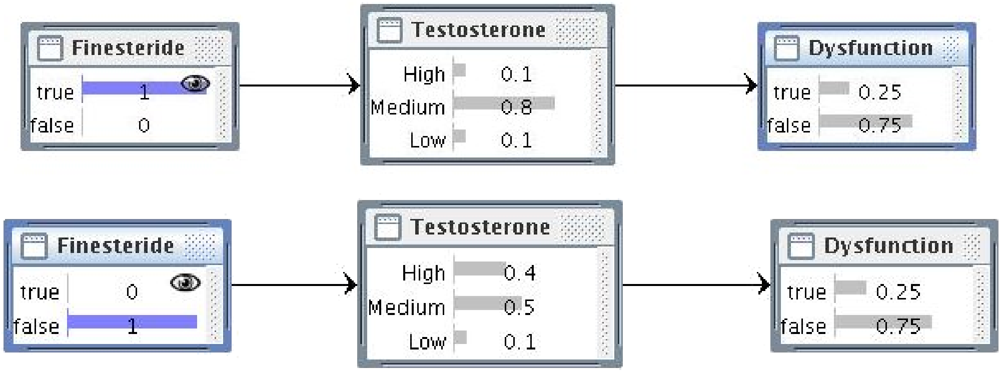

Figure 6: The finesteride example. Treatment with finesteride lowers testosterone levels. Low testosterone causes erectile dysfunction. However, finesteride does not lower testosterone sufficiently to affect erectile function.

6.2.4 Example: Finesteride and the threshold effect.

Now we present an example where noisy-or models (or any transitive theory) will not work, yet causal information provides an answer in line with intuitions.

In Section 3.1, we described a case with two causal processes: lowered testosterone levels leading to erectile dysfunction, and finesteride reducing testosterone levels. Transitivity thus dictates that prescribing finesteride will increase the chances of erectile dysfunction. However, in this case, it is just not so. No amount of finesteride can reduce testosterone to the degree needed to influence erectile dysfunction.

Figure 6 shows the scenario. The causal information values are:

- $CI(f,T) = 0.157$
- $CI(T,d) = 0.304$
- $CP(f,d) = 0.0$

Since any change to the distribution over F has no effect on the distribution over D, finesteride is correctly reported has having no causal power to bring about $D = d$ (or any other value for D).

The above examples show some of the uses for causal information in Bayesian networks. Causal information subsumes models such as Cheng's, providing similar results when both are applicable. However, causal information is applicable to a much broader range of problems, since it does not require transitivity or limiting independence assumptions.

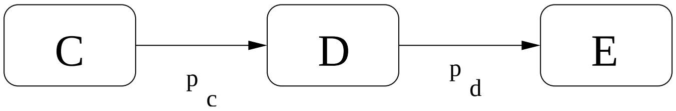

Figure 7: Two links in a noisy-or chain.

6.3 Parallel and Chained Noisy-or Networks

Now we will develop some interesting properties of noisy-or networks in particular. In our treatment we disregard root causes other than the one being analysed. We also simplify by assuming that only the causes under consideration are true parents (e.g., d is the only true parent of e in Figure 7). Our results readily generalize to other cases.

Theorem 2 *The total causal power of a noisy-or chain is the product of the powers of the individual links.*

Proof. We show that the power of a chain of two links (see Figure 7) is the product of their individual powers; it then follows that a chain can be analysed recursively.

The joint probability distribution for the network in Figure 7 is:

$$p(E,D,C) = p(E|D)p(D|C)p(C) \tag{53}$$

Therefore, assuming $p(c) \neq 0$,

$$p(e,D|c) = p(e|D)p(D|c) \tag{54}$$

$$p(e|c) = p(e|d)p(d|c) + p(e|\neg d)p(\neg d|c) \tag{55}$$

Substituting the noisy-or formula (48) gives:

$$p(e|c) = p_{de}p_{cd} + 0 \tag{56}$$

as required. □

We now show that parallel non-interactive paths are additive by using the inclusion-exclusion principle (Comtet, 1974). We refer to this special addition as 'IE-additive' and denote it by the operator $\oplus$. For two paths with powers p and q, $p \oplus q$ is defined as $p+q-pq$. The general definition is:

Definition 5

$$\oplus_i p_i = \sum_{I \in \{1,\ldots,n\}^2} (-1)^{\text{even}(|I|)} \prod_{i \in I} p_i$$

where I is a subset of the power set of indices of the p_i, $|I|$ is its cardinality and even$(\cdot)$ is 1 if its argument is an even number and zero otherwise.

Theorem 3 *The causal power of a set of parallel noisy-or chains is IE-*additive*. That is, if c is connected to e by n distinct paths, then the total power is $p_1 \oplus p_2 \oplus \ldots \oplus p_n$.*

Proof. We show that the powers of two parallel paths p_1 and p_2 are IE-additive; it immediately follows that this generalizes to n paths.

Assume then that C and E are related as in Figure 8.

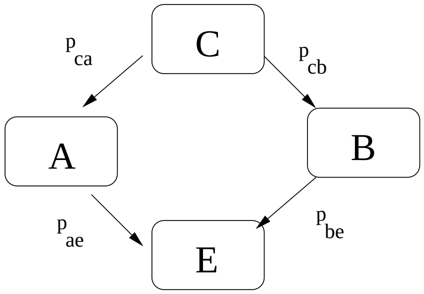

Figure 8: Two parallel paths: c is connected to e by both a and b.

Then,

(57)
$$p(E, A, B, C) = p(E|A, V)p(A, B|C)p(C)$$

Assuming $p(c) \neq 0$,

$$p(e|c) = p(e|a, b)p(a, b|c) + p(e|a, \overline{b})p(a, \overline{b}|c) + p(e|\overline{a}, b)p(\overline{a}, b|c) + p(e|\overline{a}, \overline{b})p(\overline{a}, \overline{b}|c) \tag{58}$$

Substituting the noisy-or formula gives:

$$p(e|c) = \big[1 - (1 - p_{ae})(1 - p_{be})\big]p_{ca}p_{cb} + p_{ae}p_{ca}(1 - p_{cb}) + p_{be}(1 - p_{ca})p_{cb} \tag{59}$$
$$= p_{ae}p_{ca} + p_{be}p_{cb} - p_{ca}p_{cb}p_{ae}p_{be} \tag{60}$$
$$= p_{ca}p_{ae} \oplus p_{cb}p_{be} \tag{61}$$

as required.

A or B may here represent a collection of chains, rather than individual chains, so it is obvious that IE-additivity generlizes to any number of chains relating C and E. □

7 Conclusion

We submit causal information as a significant improvement upon prior metrics of causal power. As it is based upon the information-theoretic properties of Bayesian networks, it is automatically as general as Bayesian networks, subsuming prior accounts which were based upon linear and noisy-or networks. In consequence, causal information can cope with non-transitive and interactive causes. Furthermore, our metrics directly apply to both individual variables and to sets of variables, and they allow for examination of causal powers under any context, by setting the corresponding network nodes as observed.

We have shown the relation between causal information and other information-theoretic concepts, namely mutual information and Kullback-Leibler divergence. And we have illustrated how causal information handles the cases which prior theories can cope with, as well as, in principle, any causal power analysis that might be required of the entire range of causal Bayesian networks. The most likely immediate application of our causal power metric is to enhance Bayesian network user interfaces to allow users to interrogate a network, hypothetically or retroactively, about what causal events have most influenced one or another outcome. Mutual information simpliciter cannot answer such questions, but can only inform what events are most closely related by a network probabilitically. Causal information will allow users to identify which events are most causally important for which others, and also, employing the idea of §5.4, to identify which interventions will be most effective for some desired outcome.

References

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review 104*(2), 367–405.

Comtet, L. (1974). *Advanced Combinatorics*, Chapter 4, pp. 176–178. D. Reidel Publishing Company.

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory.* John Wiley & sons.

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology.* MIT: MIT Press.

Good, I. (1961). A causal calculus. *British Journal for the Philosophy of Science 11*, 305–318.

Halpern, J. Y. and J. Pearl (2001). Causes and explanations: A structural-model approach — Part I: Causes. In J. Breese and D. Koller (Eds.), *Uncertainty in AI*, pp. 194–202.

Hiddleston, E. (2005). Causal powers. *British Journal for the Philosophy of Science 56*, 27–59.

Hitchcock, C. R. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy XCVIII*(6), 273–299.

Korb, K. B., L. R. Hope, A. E. Nicholson, and K. Axnick (2004). Varieties of causal intervention. In *PRICAI'04 – Proceedings of the 8th Pacific Rim International Conference on Artificial Intelligence*, Auckland, New Zealand, pp. 322–331.

Korb, K. B. and A. E. Nicholson (2004). *Bayesian Artificial Intelligence.* Computer Science and Data Analysis. Boca Raton: Chapman & Hall / CRC.

Neapolitan, R. E. (2003). Stochastic causality. In *International Conference on Cognitive Science, Sydney, Australia.*

Neapolitan, R. E. (2004). *Learning Bayesian Networks.* Prentice-Hall.

Novick, L. R. and P. W. Cheng (2004). Assessing interactive causal influence. *Psychological Review. 111*(2), 455–485.

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

Pearl, J. (2000a). *Causality.* New York: Cambridge.

Pearl, J. (2000b). *Causality: models, reasoning and inference.* Cambridge, UK: Cambridge University Press.

Rescorla, R. A. and A. R. Wagner (1972). A theory of Pavlovian conditioning. In A. H. Black and W. Prokasy (Eds.), *Classical Conditioning II: Current Theory and Research*, pp. 64–99. Appleton-Century-Crofts.

Salmon, W. (1980). Probabilistic causality. *Pacific Phil Qtly 61*, 50–74.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, prediction and search: second edition.* Cambridge, Massachusetts: The MIT Press.

Suppes, P. (1970). *A Probabilistic Theory of Causality.* Amsterdam.

Twardy, C. R. and K. B. Korb (2004). A criterion of probabilistic causality. *Philosophy of Science 71*, 241–62.

Verma, T. and J. Pearl (1990). Equivalence and synthesis of causal models. In *Proceedings of the sixth conference on uncertainty in artificial intelligence*, San Francisco, pp. 462–470. UAI: Morgan Kaufmann.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics 5*, 161–215.