# An Information-theoretic Causal Power Theory

Lucas R. Hope and Kevin B. Korb

School of Information Technology
Monash University
Clayton, Victoria 3800, Australia
{lhope,korb}@csse.monash.edu.au

**Abstract.** A metric of causal power can assist in developing and using causal Bayesian networks. We introduce a metric based upon information theory. 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.

## 1 Introduction

The causal interpretation of Bayesian networks has risen greatly in prominence since the development of causal discovery algorithms (Verma and Pearl, 1990; Spirtes et al., 2000; Neapolitan, 2004). However, the causal interpretation brings with it a host of difficulties, philosophical and technical, leading to various current research efforts, such as the attempt to couple the philosophical theories of probabilistic causality with causal Bayesian networks (e.g., Halpern and Pearl, 2001; Twardy and Korb, 2004).

Another long-standing research problem in philosophy and psychology has been to develop a formal theory of causal power. As causation comes in degrees, causal explanatory power — the normative attribution of an effect to one of its causes — ought also to come in degrees.

The development of a well-founded metric of causal power promises to be of wide interest: as a normative standard for assessing causal attributions; as an aid in designing Bayesian networks, by providing guidance in the interpretation of prototypes; for understanding and using probabilistic expert systems; 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, from I.J. Good (1961) to Cheng (1997), Glymour (2001) and Hiddleston (2005). A problem common to all of these theories is that they find their inspiration in simple linear (or additive) models of causality. Whereas simplicity can be an asset in developing a theory, it can be an impediment when attempting to generalize; this 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, we offer an information-theoretic metric of causal power applicable to non-linear Bayesian networks, while also illustrating their application to linear models.

## 2 A History of Causal Power

### 2.1 Good's Causal Calculus

The first serious attempt to provide a formal theory of causal power is that of I.J. Good (1961). Good's formulation 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: $e^{-R} + e^{-Q} = 1$.

Good's definition of causal strength for a direct causal link is $Q_{link}(E : C) = -\log \frac{1-p}{1-q}$,[1] where $p = P(e|c)$ and $q = P(e|\neg c)$. Good stipulates that $Q_{link}$ be non-negative, so 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 $c$ didn't happen."

The causal strength along a chain can be calculated by calculating total resistance and then converting this to causal strength:

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

where $c$ and $e$ are connected by a chain of $n$ links indexed by $i$.

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 is committed to the transitivity of causation, 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 (2004). Finesteride reduces testosterone levels (at least in rats); lowered testosterone levels can lead to erectile dysfunction. However, finesteride fails to reduce testosterone levels *sufficiently* for the follow-on erectile dysfunction to occur. 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.

### 2.2 Cheng's Power PC theory

The starting point for the probabilistic theory of casuality is probabilistic contrast: $\Delta P_c = P(e|c) - P(e|\neg c)$.[2] In this case $c$ is only a *prima facie* cause, since a common ancestor may be responsible for correlating two effects. Cheng's causal power theory attempts to overcome the limitations of *prima facie* causation.

**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 (grouped as $a$). Further, the occurence of $c$ must itself be independent of $a$. This implies that either $a$ and $c$ occur independently, or else that all the causes of $a$ are fixed.

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

[2] Suppes (1970) describes this as *prima facie causation.*

Cheng then defines the theoretical entities $p_c$ and $p_a$, respectively the causal powers of $c$ and $a$ to bring about $e$. The causal power of $c$ for $e$ is defined as the probability that $c$ produces (or generates) $e$. Since under Cheng's assumptions $e$ comes about either via $c$ or via $a$, and nothing else, this leads to:

$$P(e) = P(c)p_c + P(a)p_a - P(c)P(a)p_c p_a \tag{2}$$

(2) is used to calculate $\Delta P_c$ and then solved for $p_c$ using the above assumptions to eliminate $P(a)$ and $p_a$, giving:

$$p_c = \frac{\Delta P_c}{1 - P(e|\neg c)} \tag{3}$$

Cheng claims that this is an improvement on prior theory, such as Rescorla and Wagner (1972). Among other reasons, this is because it provides the 'correct' answer when $e$ always occurs. If $e$ always occurs, then $p_c$ is undefined, rather than zero, as Rascorla and Wagner suggested. Undefined is supposedly correct because we should be unable to assess the causes of a universal event.

**Preventative Causes** Cheng stipulates the same restrictive assumptions for preventative as for generative causes; the definitions are unchanged, except that $p_c$ is labeled preventative, leaving $a$ to be the only generative cause. Cheng says $e$ is the combination of $e$ produced by $a$ with $e$ *not being stopped* by $c$, and so:

$$P(e) = P(a)p_a(1 - P(c)p_c) \tag{4}$$

This assumes that $e$ being produced by $a$ is independent of $e$ being prevented by $c$, a rather strange assumption, as noted by Hiddleston (2005).

As with generative causes, (4) is used to find $\Delta P_c$ and then solved for $p_c$:

$$p_c = \frac{-\Delta P_c}{P(e|\neg c)} \tag{5}$$

Analogously, this leaves preventative power for an impossible $e$ undefined.

**Problems** The main difficulty for Cheng's theory is that it is extremely limited in scope. It is only defined over binary variables; but worse, the independence assumptions and limits on interactions between causes guarantees a small range of applicability.

### 2.3 Hiddleston's Causal Powers

Hiddelston's analysis of causal powers is heavily influenced by Cheng's account (Hiddleston, 2005). However, he disagrees with Cheng's formulation of preventative causes. Recall Cheng's formula (4) for how $e$ occurs when $c$ is a preventative: $P(e) = P(a)p_a(1 - P(c)p_c)$. 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 instead $P(e) = P(a)p_a(1 - P(c|a)p_{c,a})$ where $p_{c,a}$ 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 preventative barriers against some generative powers. Cheng's is a uniform barrier against all possible generative causes, while Hiddleston's only shields against a specific cause.

## 3 Causal Information

Our measure of causal power combines information theory with causal interventions on causal networks (Pearl, 2000; Korb et al., 2004).[3]

**Definition 1 Causal information** *(CI) between a cause c and an effect e in the causal model g (or, causal power of c for e) is the mutual information (MI) between the two variables in an auxiliary model $g^*$, where $g^*$ is the same as g, except the arcs between c and its parents have been cut (removed). c's distribution in $g^*$ is set as its prior in g.*

Mutual information for the discrete case is (Cover and Thomas, 1991):

$$MI(X,Y) = \sum_{x \in X, y \in Y} p(xy) \log \frac{p(xy)}{p(x)p(y)} \tag{6}$$

This has two relevant interpretations. The first is Kullback-Leibler ($KL$) divergence (or cross-entropy) between the joint probability and the product of the two marginal distributions. $KL$ divergence takes the form

$$KL(p(X), q(X)) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \tag{7}$$

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 $q$ to describe $p$. When $X$ and $Y$ are independent $p(xy) = p(x)p(y)$, so MI is the cost of assuming the two variables are independent when they may not be.

Another interpretation of mutual information is through the identity

$$MI(X,Y) = H(X) - H(X|Y)$$

The entropy $H(X)$ is the expected length of an efficient code for $X$. $H(X|Y)$ is the same, given knowledge of $Y$. So, MI information measures the aid one variable gives to the task of describing the other. However, since MI is symmetric, it cannot directly measure an asymmetric causal power. By introducing interventions, justifying the cutting of arcs in Definition 1 (e.g., Pearl, 2000; Glymour, 2001; Korb et al., 2004), causal information introduces the correct asymmetry.

There is a direct relation between causal information and KL divergence:

**Theorem 1.** *The causal information of intervention $c \in C$ wrt $E$ is:*

$$CI(C = c, E) = KL(p(E|c), p(E)) = \sum_{e \in E} p(e|c) \log \frac{p(e|c)}{p(e)} \tag{8}$$

*in auxiliary model $g^*$.*

This account has the immediate advantage of being defined in general, applying to any system for which we can find the underlying causal structure. (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 also to discrete variable models, and thus the full range of (causal) Bayesian networks, unlike any predecessors. In order to assess this account against its predecessors, however, we need to see how it applies to the simpler cases of linear and Cheng models.

[3] Space constraints force the removal of proofs to Hope and Korb (2005).

# 4 Applications

## 4.1 Path Models

In application to linear models we turn to the theory of path models, which are a general method of treating linear Gaussian models. In particular, our causal power should agree with the correlation ($r$), as calculated by the method of Wright (1934). Hope and Korb (2005) found that MI between two (unit) normals is $-\log\sqrt{1-r^2}$, where $r$ is the correlation between the two. Thus the mutual information is an increasing function of the magnitude of correlation, as we should expect and demand, since for linear models the causal information account of power is transitive, as is correlation.

## 4.2 Cheng Models

The particular feature which Cheng liked to emphasize was that her metric yielded "undefined" when the effect was impossible or necessary. Causal information is in such cases technically defined, but only because the standard convention in information theory is to treat $\log p/0$ as 0.

It is more interesting to see what causal information does with noisy-or models. Glymour (2001) noted that the assumptions Cheng applied to her models correspond to noisy-or models, which are probabilistic generalisations of the Boolean or-gate, where each parent of variable $e$ has an independant chance of triggering it, namely $p_c$ for parent $c$ (expanded to $p_{ce}$ when otherwise ambiguous). It is easiest to calculate using the probability that a cause will be inhibited: $q_i = 1 - p_i$. Let $pa(c)$ be the parents of $c$ and $pa_T(c)$ be the subset containing those which are true, then,

$$p(e|pa(e)) = 1 - \prod_{i \in pa_T(e)} q_i \tag{9}$$

The probability of $e$ being false is the probability that all the inhibitors of the occurent 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.

Now we describe some results for networks which contain only noisy-or gates. We simplify by assuming that the causes under consideration are the only true parents; the results readily generalize. (For proofs see Hope and Korb, 2005.)

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

Another result is that parallel non-interactive paths are additive, which we get by using the inclusion-exclusion principle (Comtet, 1974). We refer to this as '$IE$-addition' 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 2**

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

where $I$ is a subset of the power set of indices of the $p_i$, $|I|$ is its cardinality.

**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* $p_1 \oplus p_2 \oplus \ldots \oplus p_n$.

The causal information for Cheng models is easily derived as $CI(C = c, E) = -p_{ce} \log p(c)$, meaning that causal information is the causal power of $c$ mediated by the information content of $c$.

## 5 CONCLUSION

Causal information is far better than the metrics offered previously:

- Since it is based upon mutual information measured over Bayesian networks, it is automatically as general as Bayesian networks, including coping with interactive causes.
- The simpler properties of prior analyses, such as transitivity and additivity of causal powers, reappear when appropriate, as in linear and noisy-or models.
- As mutual information applies to individual variables or sets of variables, causal information can immediately be applied to complexes of causes.

## References


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

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

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

Glymour, C. (2001). *The Mind's Arrows*. MIT: MIT Press.

Good, I. J. (1961). A causal calculus. *Brit Jrn for Phil Sci 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.), *UAI*, pp. 194–202.

Hiddleston, E. (2005). Causal powers. *Brit Jrn for Phil Sci 56*, 27–59.

Hitchcock, C. R. (2001). The intransitivity of causation revealed in equations and graphs. *JP 98*(6), 273–299.

Hope, L. R. and K. B. Korb (2005). Information-theoretic causal power. Technical Report 2005/176, Monash University.

Korb, K. B., L. R. Hope, A. E. Nicholson, and K. Axnick (2004). Varieties of causal intervention. In *PRICAI'04*, pp. 322–331.

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

Pearl, J. (2000). *Causality*. Cambridge, UK: Cambridge.

Rescorla, R. A. and A. R. Wagner (1972). A theory of Pavlovian conditioning. In Black and Prokasy (Eds.), *Classical Conditioning II*, pp. 64–99.

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

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, prediction and search: 2nd ed.* MIT.

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 *Uncertainty in Artificial Intelligence, 6*, pp. 462–470. Morgan Kaufmann.

Wright, S. (1934). The method of path coefficients. *Ann of Math Stat 5*, 161–215.