



A refinement of the common cause principle

Nihat Ay*

Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

ARTICLE INFO

Article history:

Received 15 February 2007

Received in revised form 25 January 2008

Accepted 11 June 2008

Available online 7 September 2008

Keywords:

Causality theory

Bayesian networks

Information flows

Common cause principle

Multi-information

ABSTRACT

I study the interplay between stochastic dependence and causal relations within the setting of Bayesian networks and in terms of information theory. The application of a recently defined causal information flow measure provides a quantitative refinement of Reichenbach's common cause principle.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

1.1. The problem of system identification

Understanding the interplay and the function of a system's components generally requires not only phenomenological studies of global complex behaviour but also the study of the system's functional response to controlled experimental perturbations. Ideally, with a corresponding experimental design one aims at a complete identification of the system's mechanisms. In the context of biological systems this is clearly a problematic issue. On the one hand, a biological system may not be resistant to all experimental perturbations that would allow for its identification, and, on the other hand, the identity of the system may change as a consequence of perturbations. Furthermore, in addition to these problems, there are also technology constraints dictated by the scientific and financial means at hand. In view of these limitations, one has to address the problem of specifying the kinds of conclusions that can be drawn based on a particular set of feasible experiments. This requires a theoretical tool that is capable of modelling not only the system itself but also the data generating experimental perturbations.

Most system-theoretical models distinguish three levels of mechanistic specifications:

- (1) specification of the system's units;
- (2) structural description of the relations between the units;
- (3) functional description of the units' interactions.

Graph theory turns out to be very useful in providing a mathematical structure that serves as a model for levels (1) and (2) of the system description. In this model, the nodes of a graph represent the units of the system and the edges describe their relationship. In view of the model diversity currently present in the literature, an interpretation of edges strongly depends on the particular context. In this paper we follow Pearl's conceptual line [14] by using edges as a qualitative representation of

* Corresponding address: Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany. Fax: +49 0 341 9959.
E-mail address: nay@mis.mpg.de.

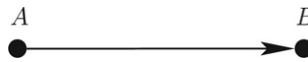


Fig. 1.

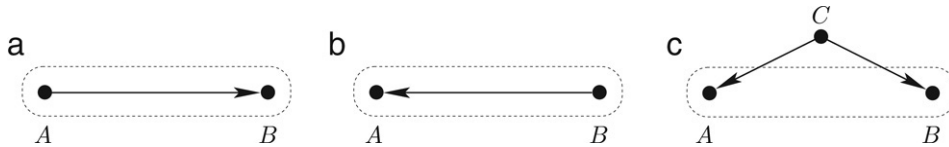


Fig. 2.

possible direct causal effects. Within Pearl's theory the functional description of level (3) is based on the so-called structural equation model originally developed in genetics [18]. In order to provide a self-consistent and coherent presentation I shall give a review of relevant concepts of Pearl's causality theory including the structural equation model (see the Appendix).

In Pearl's causality theory, the concept of *intervention*, which is intended to capture controlled experimental perturbations, plays an essential role in modelling causal effects within the system. Although experimental intervention techniques such as knockout methods in genetics are of utmost importance for understanding function in biological systems, the functional interpretation of corresponding post-interventional observations is often based on heuristic arguments and needs further justification. Here, the interplay between causal effects and general stochastic associations is still a source of confusion. In order to identify associations and the way they are generated by cause–effect relations in a given system, observation and intervention are considered as elementary experimental operations. The identification of associations of system variables is a subject of classical statistics and requires only the observation of these variables, whereas the identification of cause–effect relations in general requires experimental intervention. Pearl's theory addresses the central problem of finding particular situations in which causal effects can be identified *without* any active experimental intervention based on purely non-interventional observations. It is quite surprising that, indeed, there are several criteria, including the back-door and front-door criteria, that provide such an identifiability test for causal effects [14]. Unfortunately, the assumptions here require some structural information about the underlying network, which shows that replacement of interventional data by purely observational data is not possible without any cost.

If the structural information that is required in the above-mentioned identifiability criteria is not available, some weaker statements on the cause–effect relations can still be made. For instance, Reichenbach's principle of common cause [15], which is phrased by the slogan *no correlation without causation*, identifies a class of possible causal relations between two variables A and B if they are stochastically dependent.¹ It predicts that A is a cause of B , B is a cause of A , or there is a common cause of A and B as shown in Fig. 2. Clearly, in this qualitative statement the particular value of stochastic dependence of A and B is not used. A quantitative version of this statement would provide more information about the system. This paper extends the common cause principle in two directions: On the one hand, the main results (Theorems 3 and 4, Corollary 1) refer to more than two variables, and, on the other hand, they make *quantitative* statements by relating the stochastic dependence of these variables to causal information flows introduced in [3]. The following instructive example illustrates the main idea of the paper.

1.2. An example

Consider, as shown in Fig. 1, two units A and B , where A has a direct causal effect on B but not vice versa. In this situation, all stochastic dependence between A and B is due to the causal effect of A on B , and we can use the mutual information as a measure for the causal information flow from A to B :

$$\mathbb{F}(A \rightarrow B) := H(B) - H(B|A) = I(A : B), \quad (1)$$

where $H(B)$ denotes the entropy of B , $H(B|A)$ denotes the conditional entropy of B given A , and $I(A : B)$ is the mutual information of A and B (see Section 3). This definition is somewhat confusing. The suggestive notation of the left-hand side of (1) reflects the intuition that the causal information flow should be a directional quantity. On the other hand, mutual information is symmetric in A and B . If we just observe A and B and do not know the direction of the arrow there is no way to decide whether mutual information is an appropriate measure for the causal information flow from A to B . The situation becomes even more subtle if we have a common cause C of A and B which induces positive mutual information between A and B without having any causal information flow between A and B . Fig. 2 illustrates three completely different causal structures that allow for observational equivalence on A and B . Can we say anything about causal information flows in the network based on the observation of A and B only, without knowing the network structure? In order to illustrate in

¹ We will use the term *correlation* as the synonym of *stochastic dependence*.

what sense this is indeed possible we define the total causal information flow for the three situations as sums of the (local) information flows along the arrows as defined in (1):

$$\mathbb{F}_{(a)} := \mathbb{F}(A \rightarrow B), \quad \mathbb{F}_{(b)} := \mathbb{F}(B \rightarrow A), \quad \mathbb{F}_{(c)} := \mathbb{F}(C \rightarrow A) + \mathbb{F}(C \rightarrow B).$$

In all three cases the mutual information $I(A : B)$ provides a lower bound for the total information flow:

$$I(A : B) = \mathbb{F}_{(a)} = \mathbb{F}_{(b)} \leq \mathbb{F}_{(c)}. \tag{2}$$

The first two equalities in (2) trivially follow from the symmetry of mutual information, whereas the inequality is a consequence of the conditional independence structure of the network (c) which implies $H(A, B|C) = H(A|C) + H(B|C)$, and therefore

$$\begin{aligned} I(A : B) &= I(A : C) + I(B : C) + H(A|C) + H(B|C) - H(A, B) \\ &\leq I(A : C) + I(B : C) + H(A|C) + H(B|C) - H(A, B|C) \\ &= I(A : C) + I(B : C). \end{aligned}$$

Thus, in this example, without knowing the concrete underlying causal structure (a), (b), or (c) and corresponding mechanisms that generate the observed data distribution, we can give a lower bound for the total causal information flow in the network. This example represents a simplified quantitative version of Reichenbach’s common cause principle which, in our setting infers positive causal information flow based on positive mutual information (stochastic dependence).

1.3. Organization of the paper

In the following Section 2 some concepts of Pearl’s theory of causation are briefly outlined. This theory is based on a formal definition of intervention within the framework of Bayesian networks. Motivated by Reichenbach’s principle of common cause, in Section 3 an optimal graphical criterion for the equivalence of intervention and observation will be provided. This criterion characterizes those cases in which stochastic dependence can be interpreted as causal information flow, a notion discussed in Section 4. The main section (Section 5) provides a lower bound for information flows in terms of the multi-information of random variables and shows how this extends the common cause principle. Finally, the Appendix contains a motivation of Pearl’s concept of intervention in terms of the structural equation model, and, furthermore, the proofs of the theorems of the paper.

2. Preliminaries from Pearl’s causality theory

2.1. Directed acyclic graphs

We consider a *directed graph* $G := (V, E)$ where $V \neq \emptyset$ is a finite set of *nodes* and $E \subseteq V \times V$ is a set of *edges* between the nodes. An ordered sequence $(v_0, \dots, v_k), k \geq 0$, of distinct nodes is called a (*directed*) *path* from v_0 to v_k with *length* k if it satisfies $(v_i, v_{i+1}) \in E$ for all $i = 0, \dots, k - 1$. Given two subsets A and B of V , and a path $\gamma = (v_0, \dots, v_k)$ with $v_0 \in A$ and $v_k \in B$, we write $A \rightsquigarrow B$. If there exists a path such that $A \rightsquigarrow B$ we write $A \rightsquigarrow B$, and $A \not\rightsquigarrow B$ if this is not the case. Note that $v \rightsquigarrow v$ for all $v \in V$ (path of length 0). A *directed acyclic graph* (DAG) is a graph that does not contain two distinct nodes v_0 and v_k with $v_0 \rightsquigarrow v_k$ and $v_k \rightsquigarrow v_0$. Given a DAG, we define the *parents* of a node v as $\text{pa}(v) := \{u \in V : (u, v) \in E\}$ and its *children* as $\text{ch}(v) := \{w \in V : (v, w) \in E\}$. A set $C \subseteq V$ is called *ancestral* if for all $v \in C$ the parents $\text{pa}(v)$ are also contained in C . The smallest ancestral set that contains a set A is denoted by $\text{an}(A)$, and one has

$$\text{an}(A) = \{v \in V : v \rightsquigarrow A\}. \tag{3}$$

In his graphical models approach to causality, Pearl assumes a DAG as the structural specification of causal networks [14]. Within this specification an edge (v, w) is interpreted as a possible direct causal effect of the node v (*direct cause*) on the node w (*direct effect*). In other words, if there is no edge from v to w , then there is no possibility of directly influencing w by v . Similarly, given two non-empty and disjoint sets A and B , A is called a *cause* of B , and B an *effect* A , if $A \rightsquigarrow B$. More precisely, $A \not\rightsquigarrow B$ means that there is no possibility for direct or indirect causal influence of A on B . A node $v \in V \setminus (A \cup B)$ is called *common cause* of A and B , if there is a path from v to A that does not meet B and a path from v to B that does not meet A .

2.2. Causal effects in Bayesian networks

In addition to the structural description given by a DAG one has to specify the nodes’ interactions by a mechanistic description. In order to do so, for every node $v \in V$ we consider a finite and non-empty set \mathcal{X}_v of states. Given a subset $A \subseteq V$, we write \mathcal{X}_A instead of $\times_{v \in A} \mathcal{X}_v$ (*configuration set on A*), and we have the natural projection

$$\mathcal{X}_A : \mathcal{X}_V \rightarrow \mathcal{X}_A, \quad (x_v)_{v \in V} \mapsto x_A := (x_v)_{v \in A}.$$

Note that in the case of $A = \emptyset$ the configuration set consists of exactly one element, namely the empty configuration which we denote by ϵ .

A *distribution* on \mathcal{X}_V is a vector $p = (p(x))_x \in \mathbb{R}^{\mathcal{X}_V}$ with $p(x) \geq 0$ for all $x \in \mathcal{X}_V$ and $\sum_x p(x) = 1$. Given a distribution p on \mathcal{X}_V , the X_A 's become random variables, and we write

$$p(x_A) := \Pr\{X_A = x_A\} \quad \text{for all } x_A \in \mathcal{X}_A,$$

and, if $p(x_A) > 0$,

$$p(x_B|x_A) := \Pr\{X_B = x_B|X_A = x_A\} \quad \text{for all } x_B \in \mathcal{X}_B. \tag{4}$$

In particular, we have $p(x_B|\epsilon) = p(x_B)$ if $A = \emptyset$.

Given a DAG, we consider a family of conditional distributions $k^v(x_{pa(v)}; x_v)$, $v \in V$, that is

$$k^v(x_{pa(v)}; x_v) \geq 0 \quad \text{and} \quad \sum_{x_v} k^v(x_{pa(v)}; x_v) = 1.$$

If $pa(v) = \emptyset$ we write $k^v(x_v)$ instead of $k^v(\epsilon; x_v)$. A triple $\mathfrak{B} = (V, E, k)$ consisting of a DAG $G = (V, E)$ and such a family $k = (k^v)_{v \in V}$ of kernels is called a *Bayesian network*. Within Pearl's causality theory, the kernels k^v are interpreted mechanistically as autonomous physical processes that generate the states of the individual nodes v . This interpretation justifies assuming the stability of a node's mechanism with respect to external intervention in other nodes. The mechanistic interpretation can be motivated by the so-called *structural equation model*, which relates the kernels k^v to deterministic functions together with hidden random disturbances. This relation allows for a transparent definition of interventional operations which are essential for understanding causal effects. It turns out that all causal aspects are independent from the concrete representation of the kernels k^v by structural equations. Therefore, I continue my presentation within the context of Bayesian networks and briefly review the structural equation model in the [Appendix](#) of the paper.

The transition from the mechanistic description, given by a Bayesian network, to the phenomenological level is made by the following formula for the joint distribution $p(\mathfrak{B})$ on \mathcal{X}_V :

$$p(x) = p(\mathfrak{B}; x) := \prod_{v \in V} k^v(x_{pa(v)}; x_v). \tag{5}$$

If a given distribution p on \mathcal{X}_V can be decomposed in this way, we say that it *admits a recursive factorization according to G*. In that case one has $k^v(x_{pa(v)}; x_v) = p(x_v|x_{pa(v)})$ if $p(x_{pa(v)}) > 0$.

Given a Bayesian network $\mathfrak{B} = (V, E, k)$, one has the possibility of testing the system's reaction to external intervention. More precisely, we divide the set V into a subset A where the intervention takes place and the complement $D := V \setminus A$. Intervening in A with configuration $x'_A \in \mathcal{X}_A$ is modelled by the replacement of the mechanisms k^v , $v \in A$, by the following constant mechanisms:

$$k^v_{\text{int}}(x_{pa(v)}; x_v) := \delta_{x'_v}(x_v) = \begin{cases} 1 & \text{if } x_v = x'_v \\ 0 & \text{otherwise.} \end{cases}$$

This replacement of mechanisms leads to a new Bayesian network $\widehat{\mathfrak{B}}$ and a corresponding joint distribution according to (5) given by

$$p(x_D, x_A \parallel x'_A) := p(\widehat{\mathfrak{B}}; x) = \prod_{v \in A} \delta_{x'_v}(x_v) \prod_{v \in D} k^v(x_{pa(v)}; x_v). \tag{6}$$

Summation over all x_A finally gives us

$$\begin{aligned} p(x_D \parallel x'_A) &:= \sum_{x_A} p(x_D, x_A \parallel x'_A) \\ &= \prod_{v \in D} k^v(x_{pa(v) \setminus A}, x'_{pa(v) \cap A}; x_v). \end{aligned}$$

Replacing the pair (x_D, x'_A) by a global configuration $x = (x_D, x_A)$ allows us to write this in a more transparent way:

$$p(x_D \parallel x_A) = \prod_{v \in D} k^v(x_{pa(v)}; x_v). \tag{7}$$

Thus, compared with the pre-interventional distribution (5), the post-interventional distribution (7) is obtained simply by removing all factors $k^v(x_{pa(v)}; x_v)$ where v is an element of A (*truncated factorization*). This is the probability of observing $X_D = x_D$ after having set $X_A = x_A$. It has to be distinguished from the probability $p(x_D|x_A)$ of observing $X_D = x_D$ after having observed $X_A = x_A$. I refer to these two different ways of conditioning as *interventional* and *observational conditioning* and use two bars “||” in the first and one bar “|” in the second case. Obviously, for $A = \emptyset$ we have $p(x_B \parallel \epsilon) = p(x_B)$.

Now consider $B \subseteq D = V \setminus A$. Then

$$p(x_B \parallel x_A) = \sum_{x_{D \setminus B}} p(x_B, x_{D \setminus B} \parallel x_A) = \sum_{x_{D \setminus B}} \prod_{v \in D} k^v(x_{pa(v)}; x_v). \tag{8}$$

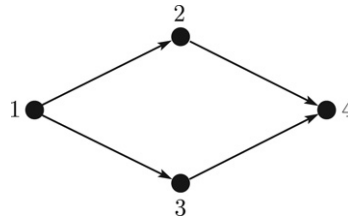


Fig. 3.

Note that interventional conditioning, in contrast to observational conditioning (4), is defined for all $x_A \in \mathcal{X}_A$. This is consistent with the semantics of a mechanism: The mechanism is defined and virtually present even for cases that do not appear in the actual distribution. The kernel $\mathcal{X}_A \times \mathcal{X}_B \rightarrow [0, 1], (x_A, x_B) \mapsto p(x_B \parallel x_A)$, is called *causal effect*. I use $p(x_B \parallel x_A)$ as a shorthand notation for the causal effect and hope that the distinction from its value at (x_A, x_B) becomes clear within the particular context.

Example 1. This instructive example illustrates the difference between interventional and observational conditioning and gives us a hint how to relate these two kinds of conditioning to each other. Consider the node set $V = \{1, 2, 3, 4\}$ and the edge set $E = \{(1, 2), (1, 3), (2, 4), (3, 4)\}$ as shown in Fig. 3. For simplicity we assume that all $k^v(x_{pa(v)}; x_v), v \in V$, are strictly positive. The joint distribution is given as

$$p(x_1, x_2, x_3, x_4) = k^1(x_1)k^2(x_1; x_2)k^3(x_1; x_3)k^4(x_2, x_3; x_4).$$

Firstly we compute the causal effect $p(x_3 \parallel x_2)$ by using formula (8):

$$\begin{aligned} p(x_3 \parallel x_2) &= \sum_{x_1, x_4} p(x_1, x_3, x_4 \parallel x_2) \\ &= \sum_{x_1, x_4} k^1(x_1)k^3(x_1; x_3)k^4(x_2, x_3; x_4) \\ &= \sum_{x_1} k^1(x_1)k^3(x_1; x_3) \\ &= p(x_3). \end{aligned}$$

As we see, this is not dependent on x_2 , which is consistent with the fact that there is no edge from 2 to 3. On the other hand, for the reason that both, node 2 and node 3 receive information from node 1, we expect that, in general, $p(x_3|x_2) \neq p(x_3)$. This can be seen as follows:

$$\begin{aligned} p(x_3|x_2) &= \frac{p(x_2, x_3)}{p(x_2)} \\ &= \frac{\sum_{x_1, x_4} k^1(x_1)k^2(x_1; x_2)k^3(x_1; x_3)k^4(x_2, x_3; x_4)}{\sum_{x_1, x_3, x_4} k^1(x_1)k^2(x_1; x_2)k^3(x_1; x_3)k^4(x_2, x_3; x_4)} \\ &= \frac{\sum_{x_1} k^1(x_1)k^2(x_1; x_2)k^3(x_1; x_3)}{\sum_{x_1} k^1(x_1)k^2(x_1; x_2)}. \end{aligned}$$

Now instead of nodes 2 and 3 we consider the nodes 1 and 4, which do not have a common cause, and compute the causal effect $p(x_4 \parallel x_1)$:

$$\begin{aligned} p(x_4 \parallel x_1) &= \sum_{x_2, x_3} p(x_2, x_3, x_4 \parallel x_1) \\ &= \sum_{x_2, x_3} k^2(x_1; x_2)k^3(x_1; x_3)k^4(x_2, x_3; x_4) \\ &= \frac{\sum_{x_2, x_3} k^1(x_1)k^2(x_1; x_2)k^3(x_1; x_3)k^4(x_2, x_3; x_4)}{k^1(x_1)} \\ &= \frac{p(x_1, x_4)}{p(x_1)} \\ &= p(x_4|x_1). \end{aligned}$$

As we see, in this case both kinds of conditioning lead to the same distribution. ★

Whether a causal effect $p(x_B \parallel x_A)$ and the corresponding conditional distribution $p(x_B|x_A)$ coincide or not depends, as **Example 1** indicates, on the presence of a common cause of A and B . In the next section, Reichenbach’s principle of common cause [15] will provide an optimal graphical condition for the equivalence of interventional and observational conditioning, which will be summarized in **Theorem 2**.

3. Entropy, mutual information, and the principle of common cause

The main intention of this paper is to carefully relate stochastic dependence (correlation) to causation in terms of information-theoretic quantities. To this end, we recall some basic definitions which already appeared in the introduction (see [6] for details): Consider two subsets A and B of V . The *entropy* of X_B is defined as

$$H_p(X_B) := - \sum_{x_B \in \mathcal{X}_B} p(x_B) \log_2 p(x_B).$$

This quantity is a natural measure of the uncertainty that one has about the outcome of X_B , that is, the information one expects to gain by observing that outcome. Knowing the outcome of X_A in general changes the uncertainty that one has about the outcome of X_B . The resulting mean uncertainty is then quantified by the *conditional entropy* of X_B given X_A :

$$H_p(X_B|X_A) := - \sum_{x_A \in \mathcal{X}_A, x_B \in \mathcal{X}_B} p(x_A, x_B) \log_2 p(x_B|x_A) \leq H_p(X_B).$$

In terms of these entropy measures, the *mutual information* of X_B and X_A is defined as

$$I_p(X_A : X_B) := H_p(X_B) - H_p(X_B|X_A) \tag{9}$$

$$= \sum_{x_A} p(x_A) \sum_{x_B} p(x_B|x_A) \log_2 \left(\frac{p(x_B|x_A)}{\sum_{x'_A} p(x'_A) p(x_B|x'_A)} \right) \tag{10}$$

$$= \sum_{x_A, x_B} p(x_A, x_B) \log_2 \left(\frac{p(x_A, x_B)}{p(x_A) p(x_B)} \right). \tag{11}$$

According to (9) it measures the uncertainty reduction of the outcome of X_B provided by the outcome of X_A and vice versa. The mutual information $I_p(X_A : X_B)$ is a natural symmetric measure for the stochastic dependence of X_A and X_B .

If not necessary, the information-theoretic quantities will be used without explicitly mentioning the underlying distribution p . For instance, $H(X)$ will be used instead of $H_p(X)$. Furthermore, basic properties of information-theoretic quantities will be applied without further explanation. A standard reference is [6].

Now we consider a Bayesian network $\mathfrak{B} = (V, E, k)$ and the ancestral sets $a := \text{an}(A)$ and $b := \text{an}(B)$ (see Eq. (3)). Then the above-mentioned quantities can be computed with respect to the joint distribution $p = p(\mathfrak{B})$, generated by \mathfrak{B} according to (5), and we obtain the following upper bounds for the mutual information of X_A and X_B .

$$\begin{aligned} I(X_A : X_B) &\leq I(X_a : X_b) \\ &= H(X_a) + H(X_b) - H(X_{a \cup b}) \\ &= H(X_{a \setminus b}, X_{a \cap b}) + H(X_{b \setminus a}, X_{a \cap b}) - H(X_{a \setminus b}, X_{a \cap b}, X_{b \setminus a}) \\ &= (H(X_{a \cap b}) + H(X_{a \setminus b}|X_{a \cap b})) + (H(X_{a \cap b}) + H(X_{b \setminus a}|X_{a \cap b})) \\ &\quad - (H(X_{a \cap b}) + H(X_{a \setminus b}|X_{a \cap b}) + H(X_{b \setminus a}|X_{a \cap b}, X_{a \setminus b})) \\ &= (H(X_{a \cap b}) + H(X_{a \setminus b}|X_{a \cap b})) + (H(X_{a \cap b}) + H(X_{b \setminus a}|X_{a \cap b})) \\ &\quad - (H(X_{a \cap b}) + H(X_{a \setminus b}|X_{a \cap b}) + H(X_{b \setminus a}|X_{a \cap b})) \\ &\quad \text{(conditional independence of } X_{a \setminus b} \text{ and } X_{b \setminus a} \text{ given } X_{a \cap b}) \\ &= H(X_{a \cap b}) \\ &\leq \sum_{v \in a \cap b} \log_2 |\mathcal{X}_v|. \end{aligned} \tag{12}$$

Clearly, if $a \cap b = \emptyset$ then the configuration set $\mathcal{X}_{a \cap b}$ consists of exactly one element which is the empty configuration. In that case the entropy vanishes and, according to (12), this implies stochastic independence of X_A and X_B . On the other hand, it is easy to see that the set $a \cap b$ is empty if and only if none of the three conditions in **Theorem 1** is satisfied, which proves the following version of Reichenbach’s principle of common cause [15].

Theorem 1 (Principle of Common Cause). Let $\mathfrak{B} = (V, E, k)$ be a Bayesian network, and let A and B be two non-empty disjoint subsets of V such that X_A and X_B are stochastically dependent with respect to the distribution $p(\mathfrak{B})$. Then one of the following conditions is satisfied:

- (1) A is a cause of B : $A \rightsquigarrow B$,
- (2) B is a cause of A : $B \rightsquigarrow A$,
- (3) A and B have a common cause: There is a node $v \in V \setminus (A \cup B)$ and a path from v to A outside of B and a path from v to B outside of A .

The principle of common cause identifies qualitative causal relations of two variables based on their stochastic dependence. The concept of d -separation, which is not applied in this paper, provides a direct proof of Theorem 1 [14,17]. The alternative proof based on inequality (12) helps in understanding the connection between the common cause principle and information theory. The elaboration of this connection is the main focus of the paper. Applying the notion of causal information flow [3], I shall provide a quantitative extension of the common cause principle which implies the estimate (12). To this end, we need a graphical criterion for the equivalence of interventional and observational conditioning where Theorem 1 can serve as a guiding scheme. Reichenbach’s principle of common cause specifies three qualitatively different but not necessarily disjoint classes of causal relations that give rise to the stochastic dependence of variables. In general, stochastic dependence is a mixed consequence of the three causal relationships (1), (2), and (3) that appear in Theorem 1. Furthermore, it is clear that in the case of (2) or (3) stochastic dependence that is not due to the causal effect of A on B is possible. Therefore, Reichenbach’s principle suggests characterizing the case where causal effects and conditional distributions coincide by assuming stochastic dependence as a consequence of causal relations of the first kind only. Therefore, we exclude the cases (2) and (3) in the following theorem.

Theorem 2. Let $\mathfrak{B} = (V, E, k)$ be a Bayesian network, let A and B be two non-empty disjoint subsets of V such that B is not a cause of A and there is no common cause of A and B . Then the conditional distribution and the causal effect coincide:

$$p(x_B \parallel x_A) = p(x_B|x_A) \quad \text{for all } x_A \text{ with positive probability } p(x_A).$$

This condition is optimal in the sense that, if it is not satisfied, then there exists a Bayesian network $\mathfrak{B}' = (V, E, k')$ for which there are x_A and x_B with $p(x_A) > 0$ and $p(x_B \parallel x_A) \neq p(x_B|x_A)$.

Theorem 2 implies that, knowing the marginal distribution $p(x_A, x_B)$ (which we assume to be strictly positive here), and knowing that B is not a cause of A and there is no common cause of A and B , one can compute the causal effect as

$$p(x_B \parallel x_A) = p(x_B|x_A) = \frac{p(x_A, x_B)}{p(x_A)} = \frac{p(x_A, x_B)}{\sum_{x'_B} p(x_A, x'_B)}. \tag{13}$$

In Pearl’s terminology, this is a special example of an *identifiable* causal effect $p(x_B \parallel x_A)$. On the other hand, if we do not know anything about the underlying graph structure, and, in particular, if we do not know whether the conditions of Theorem 2 are satisfied or not, it is not possible to use formula (13) for computing the causal effects explicitly. But this does not mean that we cannot say anything about the causal structure as the principle of common cause shows. The intention of this paper is to point out that we can say even more than that by using a quantitative extension of the common cause principle based on the notion of causal information flows [3]. This notion is introduced in the following section.

4. Causal information flows

In order to quantify causal effects instead of general associations, in [3] we suggested replacing the conditional probabilities $p(x_B|x_A)$ in (10) by the interventional probabilities $p(x_B \parallel x_A)$. This suggestion was based on concepts that had been discussed in the previous work [10,2]. Given a Bayesian network $\mathfrak{B} = (V, E, k)$ we consider the joint distribution p generated according to (5). Replacing $p(x_B|x_A)$ by $p(x_B \parallel x_A)$ means that we consider a new joint distribution $\hat{p}(x_A, x_B) := p(x_A) p(x_B \parallel x_A)$ and the corresponding mutual information of X_A and X_B :

$$\begin{aligned} \mathbb{F}_{\mathfrak{B}}(X_A \rightarrow X_B) &:= I_{\hat{p}}(X_A : X_B) \\ &= \sum_{x_A} p(x_A) \sum_{x_B} p(x_B \parallel x_A) \log_2 \left(\frac{p(x_B \parallel x_A)}{\sum_{x'_A} p(x'_A) p(x_B \parallel x'_A)} \right). \end{aligned}$$

This measure quantifies the causal effect of A on B and has been termed (*causal*) *information flow* in [3]. We also use the notation \mathbb{F} where the explicit reference to the underlying Bayesian network \mathfrak{B} is omitted.

If B is not a cause of A and there is no common cause of A and B , then, according to Theorem 2, the mutual information of X_A and X_B and the causal information flow from A to B coincide:

$$I(X_A : X_B) = \mathbb{F}(X_A \rightarrow X_B). \tag{14}$$

In the following examples, all pairs of sets A and B have this property.

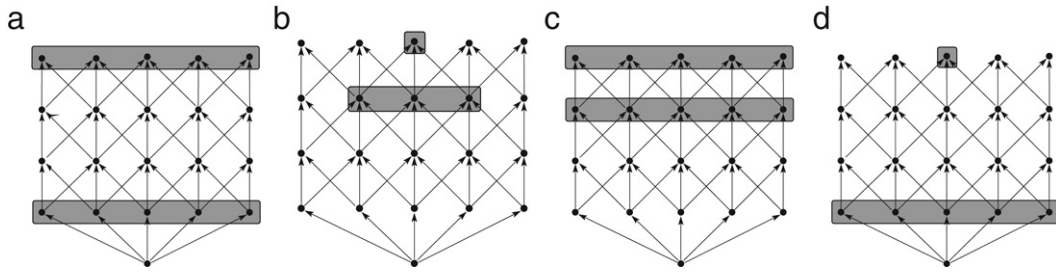


Fig. 4.

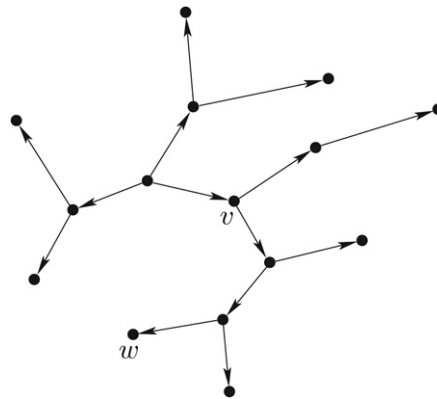


Fig. 5.

Examples 2. (1) *Direct causes.* Let $\mathfrak{B} = (V, E, k)$ be a Bayesian network. Then for all $v \in V$ the sets $A = \text{pa}(v)$ and $B = \{v\}$ satisfy

$$I(X_{\text{pa}(v)} : X_v) = \mathbb{F}(X_{\text{pa}(v)} \rightarrow X_v). \tag{15}$$

(2) *Ancestral sets.* Let $\mathfrak{B} = (V, E, k)$ be a Bayesian network, let A be an ancestral subset of V , and let B be an arbitrary subset of $V \setminus A$. Then (14) holds.

(3) *Feed-forward networks.* Here we consider a Bayesian network $\mathfrak{B} = (V, E, k)$ with a particular DAG structure, known as *feed-forward structure*: Given L non-empty, finite, and disjoint sets V_1, \dots, V_L (*layers*), we consider a node set V and an edge set E satisfying

$$V = \bigcup_{i=1}^L V_i, \quad E \subseteq \bigcup_{i=1}^{L-1} (V_i \times V_{i+1}).$$

Furthermore, let $A \subseteq V_i$ and $B \subseteq V_j$ be two non-empty sets with $i < j$. Then:

$$\text{an}(B) \cap V_i \subseteq A \Rightarrow \tag{14}. \tag{16}$$

In the networks (a)–(d) of Fig. 4, A is given by the lower subset and B is given by the upper subset. The subsets A and B of the feed-forward networks shown in Fig. 4 relate to various information-theoretic studies within theoretical neuroscience. In these studies, the maximization of a corresponding information-theoretic quantity, often causally interpreted as information flow, has been considered as a first principle of learning and could provide explanation for experimental findings [11,16,4,5,1,13]. On the other hand, it is clear that these studies are restricted to very special cases and extensions to more general situations including information flows in recurrent networks require a careful consideration of causality.

(4) *Trees.* Let $\mathfrak{B} = (V, E, k)$ be a Bayesian network where $G = (V, E)$ is a tree, which means that there are no cycles in the undirected version of G (see Fig. 5). If $v, w \in V$ are two distinct nodes that satisfy $v \rightsquigarrow w$, the connecting path is unique. This implies (14). ★

5. Stochastic dependence and information flows

5.1. Local information flows

According to Eq. (15) one can interpret the mutual information of $X_{\text{pa}(v)}$ and X_v causally as information flow. Given distinct nodes v_1, \dots, v_n in V , Eq. (15) can also be used to relate more general stochastic dependence of these nodes to causal information flows. In order to derive such a relationship, we consider the so-called *multi-information* of (discrete) random variables X_1, \dots, X_n with joint distribution p :

$$\begin{aligned} I_p(X_1, \dots, X_n) &:= \sum_{x_1, \dots, x_n} p(x_1, \dots, x_n) \log_2 \left(\frac{p(x_1, \dots, x_n)}{p(x_1) \dots p(x_n)} \right) \\ &= \sum_{i=1}^n H_p(X_i) - H_p(X_1, \dots, X_n). \end{aligned}$$

This is an extension of the mutual information to the case of more than two random variables. Now we want to address the following problem: Given the stochastic dependence of nodes $\{v_1, \dots, v_n\} \subseteq V$, which is measured by the multi-information of X_{v_1}, \dots, X_{v_n} , can we say anything about the required causal information flows in the system that lead to that stochastic dependence? In order to illustrate how this can be done, we consider the instructive special case of an ancestral set $A \subseteq V$ of a Bayesian network $\mathfrak{B} = (V, E, k)$. We choose a numbering $v_i, i \in \{1, \dots, n\}, n = |A|$, of the nodes in A that satisfies

$$v_i \in \text{pa}(v_j) \Rightarrow i < j. \tag{17}$$

Note that such a numbering always exists in an ancestral set. With the chain rule for the entropy we obtain

$$\begin{aligned} I(X_{v_1}, \dots, X_{v_n}) &= \sum_{i=1}^n H(X_{v_i}) - H(X_{v_1}, \dots, X_{v_n}) \\ &= \sum_{i=1}^n H(X_{v_i}) - \sum_{i=1}^n H(X_{v_i} | X_{v_1}, \dots, X_{v_{i-1}}) \quad (\text{entropy chain rule}) \\ &= \sum_{i=1}^n H(X_{v_i}) - \sum_{i=1}^n H(X_{v_i} | X_{\text{pa}(v_i)}) \\ &\quad (\text{conditional independence of } X_{v_i} \text{ and } X_{v_1}, \dots, X_{v_{i-1}} \text{ given } X_{\text{pa}(v_i)}) \\ &= \sum_{i=1}^n I(X_{\text{pa}(v_i)} : X_{v_i}). \end{aligned}$$

According to (15), this is equivalent to

$$I(X_{v_1}, \dots, X_{v_n}) = \sum_{i=1}^n \mathbb{F}(X_{\text{pa}(v_i)} \rightarrow X_{v_i}). \tag{18}$$

Note that this result is valid for any numbering of the nodes in A and does not depend on the particular one that we have chosen. Only the existence of a numbering that satisfies (17) is required, which is guaranteed by the assumption that A is an ancestral set.

Eq. (14), with its corresponding Examples 2, and Eq. (18) show how stochastic dependence can be expressed in terms of causal information flows. But we have to keep in mind that these equalities are based on some knowledge about the network structure. For instance, in order to obtain (18) we assumed that the set $A = \{v_1, \dots, v_n\}$ is an ancestral set. But how can we relate the stochastic dependence of the nodes in A to causal information flows if we do not assume A to be ancestral? It is somewhat surprising that we can skip this assumption and are still able to identify the stochastic dependence at least as a lower bound for the total information flow. This is the message of the following theorem.

Theorem 3. *Let $\mathfrak{B} = (V, E, k)$ be a Bayesian network, and let v_1, \dots, v_n be distinct elements of V . Then*

$$I(X_{v_1}, \dots, X_{v_n}) \leq \max_{j \in \{1, \dots, n\}} \sum_{\substack{i=1 \\ i \neq j}}^n \mathbb{F}(X_{\text{pa}(v_i)} \rightarrow X_{v_i}) \leq \sum_{i=1}^n \mathbb{F}(X_{\text{pa}(v_i)} \rightarrow X_{v_i}). \tag{19}$$

Intuitively, this theorem states that, in order to generate a particular value of stochastic dependence in $A = \{v_1, \dots, v_n\}$, the sum of local information flows within and into A has to exceed that value.

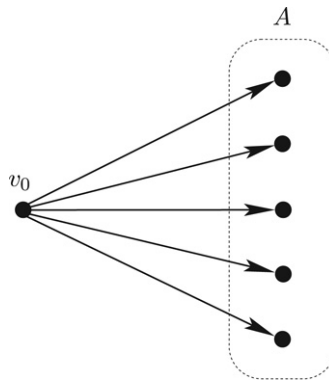


Fig. 6.

Example 3. Consider the units $V = \{v_0, v_1, \dots, v_n\}$ with state sets $\{0, 1\}$. Assume that the unit v_0 randomly generates a state X_{v_0} and transmits it without any modification to all other units. Thus, $X_v = X_{v_0}$ (almost surely) for all v . The DAG is given by the edge set $E = \{v_0\} \times \{v_1, \dots, v_n\}$ (see Fig. 6). Now consider the subset $A = \{v_1, \dots, v_n\}$. There are no direct causal effects within A . Nevertheless, in this example, the stochastic dependence of the variables $X_v, v \in A$, gives a good estimate of the total causal information flow in the network:

$$\begin{aligned} \sum_{i=1}^n \mathbb{F}(X_{\text{pa}(v_i)} \rightarrow X_{v_i}) &= nH(X_{v_0}) \\ &\geq (n - 1)H(X_{v_0}) \\ &= \max_{j \in \{1, \dots, n\}} \sum_{\substack{i=1 \\ i \neq j}}^n \mathbb{F}(X_{\text{pa}(v_i)} \rightarrow X_{v_i}) \\ &= I(X_{v_1}, \dots, X_{v_n}). \quad \star \end{aligned}$$

5.2. Information flows from common causes

Now we come back to Reichenbach’s common cause principle and its quantitative extension. Theorem 3 is not directly applicable to this end, because it relates stochastic dependence only to local information flows, whereas flows from common causes, for instance, are in general non-local and originating from farther regions of the network. On the other hand, Theorem 3 can be applied to a kind of “coarse-grained” Bayesian network explicitly defined in the proof of Theorem 4 (see the Appendix), in order to obtain a quantitative refinement of the common cause principle in terms of information flows. This refinement is the content of Theorem 4 and its Corollary 1. The formulation of these results requires some definitions which I introduce, for didactical reasons, in three steps.

Step 1: Given a Bayesian network $\mathfrak{B} = (V, E, k)$ and distinct nodes $v_1, \dots, v_n \in V$, consider the map

$$\varphi : V \rightarrow 2^{\{1, \dots, n\}}, \quad v \mapsto \varphi(v) := \{j \in \{1, \dots, n\} : v \rightsquigarrow v_j\}, \tag{20}$$

where $2^{\{1, \dots, n\}}$ denotes the power set of $\{1, \dots, n\}$. The φ -preimage of an element $S \in 2^{\{1, \dots, n\}}$, which we denote by α_S , coincides with the set of nodes $v \in V$ that satisfy $v \rightsquigarrow v_i$ if and only if $i \in S$, that is

$$\alpha_S = \left(\bigcap_{i \in S} \text{an}(v_i) \right) \cap \left(\bigcap_{i \in \{1, \dots, n\} \setminus S} (V \setminus \text{an}(v_i)) \right). \tag{21}$$

We consider the set $\mathcal{V} := \{S \subseteq \{1, \dots, n\} : \alpha_S \neq \emptyset\}$. Note that the cardinality of \mathcal{V} is upper bounded by the cardinality of V . The $\alpha_S, S \in \mathcal{V}$, are the atoms of a partition of V , and we have

$$\text{an}(v_i) = \bigcup_{\substack{S \in \mathcal{V} \\ S \supseteq \varphi(v_i)}} \alpha_S. \tag{22}$$

Step 2: Based on the decomposition (22) of the ancestral set $\text{an}(v_i)$ into atoms we consider the following bipartition into a pair of disjoint sets:

$$A_i := \bigcup_{\substack{S \in \mathcal{V} \\ S \supseteq \varphi(v_i)}} \alpha_S \quad \text{and} \quad B_i := \alpha_{\varphi(v_i)} = \text{an}(v_i) \setminus A_i. \tag{23}$$

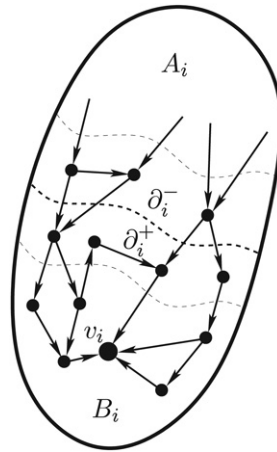


Fig. 7.

In other words, the set B_i is defined to be the atom that contains v_i , and A_i is the complement of B_i in $\text{an}(v_i)$. Obviously,

$$v_i \in B_j \iff i = j. \tag{24}$$

Otherwise one would have the contradiction that two distinct nodes v_i and v_j satisfy $v_i \rightsquigarrow v_j$ and $v_j \rightsquigarrow v_i$. We will see in the Appendix that A_i has the following explicit representation which directly implies that A_i is an ancestral set:

$$A_i = \{v \in V : v \rightsquigarrow v_i \text{ and there exists } j \neq i \text{ satisfying } v_i \not\rightsquigarrow v_j \text{ and } v \rightsquigarrow v_j\}. \tag{25}$$

Note that in the case of $v_i \not\rightsquigarrow v_j$ for all $i \neq j$ the individual intersections $A_i \cap A_j$, $i \neq j$, consist of the common causes of v_i and v_j . This interesting special case is considered in Corollary 1.

Step 3: Theorem 4 provides an upper bound of the multi-information in terms of information flows from the A_i 's to the B_i 's. These flow values do not change if we replace the cause and effect sets by smaller sets ∂_i^- and ∂_i^+ at the “boundary” between A_i and B_i as illustrated in Fig. 7. In order to define the sets ∂_i^- and ∂_i^+ , we need the notion of the so-called Markov blanket of a node v (see [7], page 71). It consists of the parents, the children, and the children’s parents of v :

$$\text{bl}(v) := \text{pa}(v) \cup \text{ch}(v) \cup \{w \in V : \text{ch}(w) \cap \text{ch}(v) \neq \emptyset\}.$$

A Markov blanket of a set A is defined as

$$\text{bl}(A) := \bigcup_{v \in A} (\text{bl}(v) \setminus A).$$

We apply this definition to the sets A_i and B_i with respect to the subgraph G_{a_i} induced by the ancestral set $a_i := \text{an}(v_i)$, that is $G_{a_i} = (a_i, E \cap (a_i \times a_i))$:

$$\partial_i^- := \text{bl}(B_i) \quad \text{and} \quad \partial_i^+ := \text{bl}(A_i). \tag{26}$$

The fact that A_i is an ancestral set implies that there is no edge starting in B_i and ending in A_i . Therefore we have

$$\partial_i^- = \{v \in A_i : \text{ch}(v) \cap B_i \neq \emptyset\} \quad \text{and} \tag{27}$$

$$\partial_i^+ = \{w \in B_i : \text{pa}(w) \cap A_i \neq \emptyset, \text{ or there exists } v \in A_i \text{ with } \text{ch}(v) \cap \text{ch}(w) \neq \emptyset\}. \tag{28}$$

In order to have a better intuitive understanding of Steps 1–3, we apply the corresponding definitions to the following simple and very special example.

Example 4. Consider the nodes $v_1 = 8$, $v_2 = 11$, and $v_3 = 14$ of the tree in Fig. 8. The subtree that is emphasized in Fig. 8 by solid lines and dots, is given by the ancestral node set $\text{an}(\{8, 11, 14\})$, which is the union of

$$\text{an}(8) = \{1, 2, 4, 8\}, \quad \text{an}(11) = \{1, 2, 5, 11\}, \quad \text{and} \quad \text{an}(14) = \{1, 3, 7, 14\}.$$

These ancestral sets generate a partition of the node set consisting of the following atoms (see representation (21)):

$$\begin{aligned} \alpha_\emptyset &= \{6, 9, 10, 12, 13, 15\}, & \alpha_{\{1,2\}} &= \{2\}, & \alpha_{\{1,2,3\}} &= \{1\}, \\ \alpha_{\{1\}} &= \{4, 8\}, & \alpha_{\{2\}} &= \{5, 11\}, & \alpha_{\{3\}} &= \{3, 7, 14\}. \end{aligned}$$

Now we divide each ancestral set $\text{an}(v_i)$ into a disjoint cause–effect pair A_i and B_i . The effect set B_i is given by the atom that contains v_i , and the cause set A_i is simply the complement of B_i in the ancestral set of v_i . From Fig. 8 we can easily read off

$$A_1 = \{1, 2\}, B_1 = \{4, 8\}, \quad A_2 = \{1, 2\}, B_2 = \{5, 11\}, \quad \text{and} \quad A_3 = \{1\}, B_3 = \{3, 7, 14\}.$$

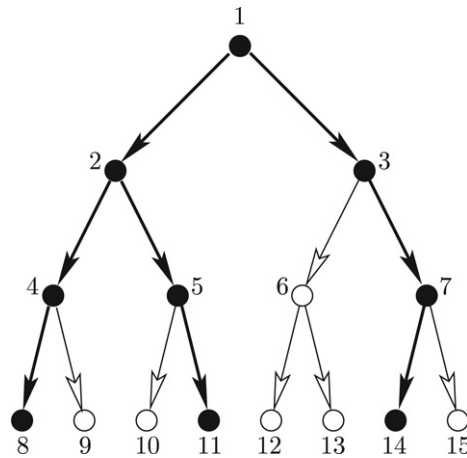


Fig. 8.

Replacing each A_i by the smaller cause set $\partial_i^- \subseteq A_i$ and each B_i by the smaller effect set $\partial_i^+ \subseteq B_i$ (see Eqs. (27) and (28)) we obtain

$$\partial_1^- = \{2\}, \partial_1^+ = \{4\}, \quad \partial_2^- = \{2\}, \partial_2^+ = \{5\}, \quad \text{and} \quad \partial_3^- = \{1\}, \partial_3^+ = \{3\}. \quad \star$$

We are now ready for the quantitative refinement of the common cause principle in terms of information flows.

Theorem 4. Let $\mathfrak{B} = (V, E, k)$ be a Bayesian network, and let v_1, \dots, v_n be distinct elements of V . Then for all i we have

$$\mathbb{F}(X_{A_i} \rightarrow X_{B_i}) = I(X_{A_i} : X_{B_i}) = I(X_{\partial_i^-} : X_{\partial_i^+}) = \mathbb{F}(X_{\partial_i^-} \rightarrow X_{\partial_i^+}), \tag{29}$$

and the following inequalities hold:

$$I(X_{v_1}, \dots, X_{v_n}) \leq I(X_{B_1}, \dots, X_{B_n}) \tag{30}$$

$$\leq \max_{j \in \{1, \dots, n\}} \sum_{\substack{i=1 \\ i \neq j}}^n \mathbb{F}(X_{A_i} \rightarrow X_{B_i}) \leq \sum_{i=1}^n \mathbb{F}(X_{A_i} \rightarrow X_{B_i}). \tag{31}$$

Note that the left-hand side of (30) is a function of the (observed) joint distribution and does not explicitly depend on the network structure, whereas the computation of the individual information flow terms of both sums of (31) does require explicit knowledge about the network structure.

In the case where the observed nodes v_1, \dots, v_n do not influence each other, that is $v_i \not\rightsquigarrow v_j$ for all $i \neq j$, one has

$$A_i = \{v \in V : v \rightsquigarrow v_i \text{ and there exists } j \neq i \text{ satisfying } v \rightsquigarrow v_j\} = \bigcup_{j \neq i} (A_i \cap A_j).$$

Here, as stated directly after Eq. (25), the intersections $A_i \cap A_j$, $i \neq j$, consist of the common causes of v_i and v_j . Therefore, in this case, the application of Theorem 4 provides an upper bound of the multi-information of the observed nodes in terms of information flows from their common causes.

Corollary 1. If, in the situation of Theorem 4, $v_i \not\rightsquigarrow v_j$ holds for all $i \neq j$, then

$$\mathbb{F}(X_{A_i} \rightarrow X_{v_i}) = I(X_{A_i} : X_{v_i}) = I(X_{\partial_i^-} : X_{v_i}) = \mathbb{F}(X_{\partial_i^-} \rightarrow X_{v_i}), \tag{32}$$

and

$$I(X_{v_1}, \dots, X_{v_n}) \leq \max_{j \in \{1, \dots, n\}} \sum_{\substack{i=1 \\ i \neq j}}^n \mathbb{F}(X_{A_i} \rightarrow X_{v_i}) \leq \sum_{i=1}^n \mathbb{F}(X_{A_i} \rightarrow X_{v_i}). \tag{33}$$

Obviously, in the situation of Corollary 1, the inequalities (33) are sharper than the corresponding inequalities (31). On the other hand, it is important to keep in mind that, in contrast to the situation of Theorem 4, not only the individual information flow terms of the inequalities (33) but even the applicability of these inequalities requires some knowledge about the network structure, namely that $v_i \not\rightsquigarrow v_j$ for all $i \neq j$. In other words, if we do not know anything about the

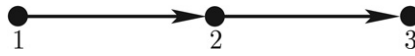


Fig. 9.

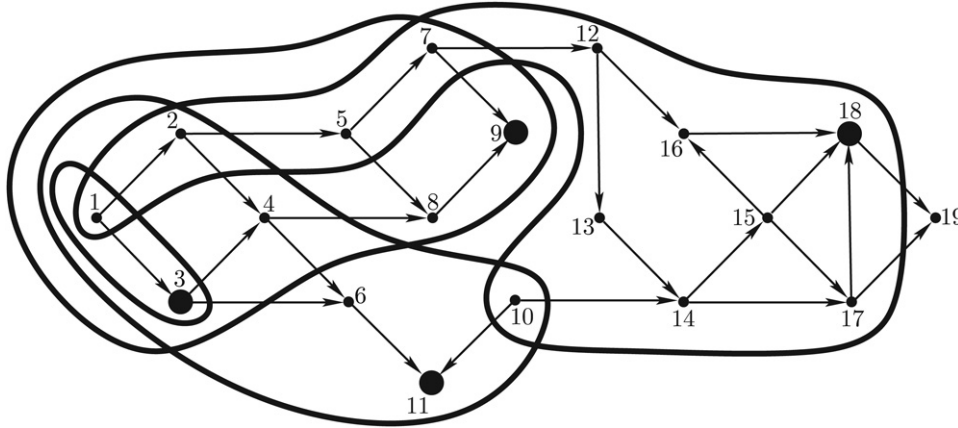


Fig. 10.

underlying network we cannot say whether (33) is true, whereas the validity of (31) is guaranteed without structural knowledge although the computation of the individual information flow terms *does* require such knowledge. This constitutes the conceptual difference between Theorem 4 and Corollary 1.

Examples 5. (1) *Data processing inequality.* Consider the the simple DAG shown in Fig. 9. If we observe the first node $v_1 = 1$ and third node $v_2 = 3$, we have

$$A_1 = \emptyset, B_1 = \{1\}, \quad \partial_1^- = \emptyset, \partial_1^+ = \emptyset,$$

$$A_2 = \{1\}, B_2 = \{2, 3\}, \quad \partial_2^- = \{1\}, \partial_2^+ = \{2\}.$$

Then (29) and (31) imply the inequality

$$I(X_1 : X_3) \leq \mathbb{F}(X_1 \rightarrow X_2) = I(X_1 : X_2).$$

This is well known as *data processing inequality* (Theorem 2.8.1 of [6]).

(2) *Example 4 continued.* Here we continue Example 4 by applying Theorem 4. The second inequality of (31) implies

$$I(X_8, X_{11}, X_{14}) \leq \mathbb{F}(X_{\{1,2\}} \rightarrow X_{\{4,8\}}) + \mathbb{F}(X_{\{1,2\}} \rightarrow X_{\{5,11\}}) + \mathbb{F}(X_1 \rightarrow X_{\{3,7,14\}}). \tag{34}$$

According to (29), we can remove some of the cause and effect nodes without changing the information flows on the right-hand side of (34):

$$I(X_8, X_{11}, X_{14}) \leq \mathbb{F}(X_2 \rightarrow X_4) + \mathbb{F}(X_2 \rightarrow X_5) + \mathbb{F}(X_1 \rightarrow X_3). \tag{35}$$

The additional structural knowledge that there are no edges between the nodes 8, 11, and 14, allows us to apply Corollary 1 and thereby further sharpen the upper bound of the multi-information:

$$I(X_8, X_{11}, X_{14}) \leq \mathbb{F}(X_2 \rightarrow X_8) + \mathbb{F}(X_2 \rightarrow X_{11}) + \mathbb{F}(X_1 \rightarrow X_{14}). \tag{36}$$

This is a refinement of the estimate (35).

(3) *More “entangled” causal relations.* This example illustrates how Theorem 4 works in situations where the causal relations of the nodes are more “entangled” than those of the tree in Example 4. Consider the DAG shown in Fig. 10. We assume that we observe $v_1 = 3, v_2 = 9, v_3 = 11$, and $v_4 = 18$ (these nodes are emphasized by bigger dots). The corresponding ancestral sets $an(v_i), i = 1, 2, 3, 4$, are encircled by individual closed lines. From Fig. 10 we can directly read off the individual cause–effect sets of Theorem 4:

$$A_1 = \{1\}, B_1 = \{3\}, \quad \partial_1^- = \{1\}, \partial_1^+ = \{3\},$$

$$A_2 = \{1, 2, 3, 4, 5, 7\}, B_2 = \{8, 9\}, \quad \partial_2^- = \{4, 5, 7\}, \partial_2^+ = \{8, 9\},$$

$$A_3 = \{1, 2, 3, 4, 10\}, B_3 = \{6, 11\}, \quad \partial_3^- = \{3, 4, 10\}, \partial_3^+ = \{6, 11\},$$

$$A_4 = \{1, 2, 5, 7, 10\}, B_4 = \{12, 13, 14, 15, 16, 17, 18\}, \quad \partial_4^- = \{7, 10\}, \partial_4^+ = \{12, 13, 14\}.$$

Using the reduced sets ∂_i^- and ∂_i^+ , Theorem 4 implies

$$I(X_3, X_9, X_{11}, X_{18}) \leq \mathbb{F}(X_1 \rightarrow X_3) + \mathbb{F}(X_{\{4,5,7\}} \rightarrow X_{\{8,9\}}) + \mathbb{F}(X_{\{3,4,10\}} \rightarrow X_{\{6,11\}}) + \mathbb{F}(X_{\{7,10\}} \rightarrow X_{\{12,13,14\}}). \quad \star$$

In Section 3 we have seen that the inequality (12) implies the common cause principle (Theorem 1). As a concluding remark I apply the first estimate of (31) to two nodes v_1 and v_2 and thereby identify it as a refinement of (12) in the case where A and B have cardinality one. With $a := \text{an}(v_1)$ and $b := \text{an}(v_2)$ there are three qualitatively different cases that determine the sets A_i and B_i used in Theorem 4:

	A_1	A_2	B_1	B_2
Case 1: $v_1 \in a \setminus b, v_2 \in b \setminus a$	$a \cap b$	$a \cap b$	$a \setminus b$	$b \setminus a$
Case 2: $v_1 \in a \setminus b, v_2 \in a \cap b$	$a \cap b$	\emptyset	$a \setminus b$	$a \cap b$
Case 3: $v_1 \in a \cap b, v_2 \in b \setminus a$	\emptyset	$a \cap b$	$a \cap b$	$b \setminus a$

Straightforward application of Theorem 4 gives us the following inequality (37):

$$I(X_{v_1} : X_{v_2}) \leq \begin{cases} \max \{ \mathbb{F}(X_{a \cap b} \rightarrow X_{a \setminus b}), \mathbb{F}(X_{a \cap b} \rightarrow X_{b \setminus a}) \} & \text{in case 1} \\ \mathbb{F}(X_{a \cap b} \rightarrow X_{a \setminus b}) & \text{in case 2} \\ \mathbb{F}(X_{a \cap b} \rightarrow X_{b \setminus a}) & \text{in case 3} \end{cases} \quad (37)$$

$$\begin{aligned} &\leq \max \{ \mathbb{F}(X_{a \cap b} \rightarrow X_{a \setminus b}), \mathbb{F}(X_{a \cap b} \rightarrow X_{b \setminus a}) \} & (38) \\ &= \max \{ I(X_{a \cap b} : X_{a \setminus b}), I(X_{a \cap b} : X_{b \setminus a}) \} \quad (\text{Examples 2 (2)}) \\ &\leq H(X_{a \cap b}) \\ &\leq \sum_{v \in a \cap b} \log_2 |\mathcal{X}_v|. \end{aligned}$$

As we see, (37) and (38) are refinements of the previous estimate (12) in the case where A and B are sets of size one. In particular, they also imply the common cause principle of Theorem 1 in that case.

Acknowledgments

I sincerely thank Daniel Polani for his comments and suggestions which were not only extremely helpful but in many respects crucial and thereby greatly improved this paper. I am also grateful for many stimulating discussions with Shun-ichi Amari, Nils Bertschinger, Jürgen Jost, David Krakauer, Eckehard Olbrich, and Bastian Steudel and for the referees' professional work. Furthermore, I would like to thank the Brain Science Institute at RIKEN for hosting me during the initial work on this paper and the Santa Fe Institute for supporting me as an external professor.

Appendix

A.1. Structural equation model and intervention

In Section 2.2 we used Bayesian networks as a formal description of causal interactions. Although it turns out that this is a sufficient model for understanding causal effects, for some scientists, including Pearl, it appears more intuitive to assume a deterministic nature of functional mechanisms. Following Pearl, the mechanisms of the nodes $v \in V$ are described by distributions (disturbances) d_v on sets \mathcal{U}_v , and deterministic maps $f_v : \mathcal{X}_{\text{pa}(v)} \times \mathcal{U}_v \rightarrow \mathcal{X}_v$. The corresponding equations

$$x_v = f_v(x_{\text{pa}(v)}, u_v), \quad v \in V, \quad (39)$$

are called *structural equations* [18,9,8]. The disturbances are assumed to be mutually independent. A DAG $G = (V, E)$ together with a family of disturbances d_v and maps $f_v, v \in V$, is called a *causal model*. A causal model \mathcal{C} defines the following joint distribution on $\mathcal{X}_V \times \mathcal{U}_V$:

$$p(\mathcal{C}; x, u) = \prod_{v \in V} d_v(u_v) \delta_{f_v(x_{\text{pa}(v)}, u_v)}(x_v), \quad x \in \mathcal{X}_V, u \in \mathcal{U}_V. \quad (40)$$

If we take the marginal of that distribution, we obtain a distribution on \mathcal{X}_V :

$$\begin{aligned} p(\mathcal{C}; x) &= \sum_u p(\mathcal{C}; x, u) \\ &= \sum_u \prod_{v \in V} d_v(u_v) \delta_{f_v(x_{\text{pa}(v)}, u_v)}(x_v) \end{aligned}$$

$$\begin{aligned}
 &= \prod_{v \in V} \left(\sum_{u_v} d_v(u_v) \delta_{f_v(x_{pa(v)}, u_v)}(x_v) \right) \\
 &= \prod_{v \in V} k^v(x_{pa(v)}; x_v),
 \end{aligned}$$

with

$$k^v(x_{pa(v)}; x_v) := \sum_{u_v} d_v(u_v) \delta_{f_v(x_{pa(v)}, u_v)}(x_v).$$

These are exactly the kernels that we already considered in Bayesian networks.

In order to describe interventions in a causal model we split the node set V into a subset A of nodes that are intervened and the subset $D := V \setminus A$ of remaining nodes which are observed. Let x'_A be a configuration on A . Setting X_A to x'_A means replacing all mechanisms f_v , $v \in A$, by constants. Then we have to replace the structural equation (39) as follows:

$$\begin{aligned}
 x_v &= f_v(x_{pa(v) \setminus A}, x'_{pa(v) \cap A}, u_v), \quad v \in V \setminus A, \\
 x_v &= x'_A, \quad v \in A.
 \end{aligned}$$

This gives the new causal model $\widehat{\mathcal{C}}$ where the maps f_v are replaced by the new maps $\widehat{f}_v := x'_v$ if $v \in A$, and $\widehat{f}_v := f_v$ if $v \notin A$. This new causal model defines the following joint distribution which is a modification of (40):

$$p(\widehat{\mathcal{C}}; x, u) := \prod_{v \in A} d_v(u_v) \delta_{x'_v}(x_v) \prod_{v \in V \setminus A} d_v(u_v) \delta_{f_v(x_{pa(v) \setminus A}, x'_{pa(v) \cap A}, u_v)}(x_v).$$

This implies

$$\begin{aligned}
 p(\widehat{\mathcal{C}}; x_A, x_D) &= \sum_u p(\widehat{\mathcal{C}}; x_A, x_D, u) \\
 &= \sum_u \prod_{v \in A} d_v(u_v) \delta_{x'_v}(x_v) \prod_{v \in V \setminus A} d_v(u_v) \delta_{f_v(x_{pa(v) \setminus A}, x'_{pa(v) \cap A}, u_v)}(x_v) \\
 &= \left(\prod_{v \in A} \sum_{u_v} d_v(u_v) \delta_{x'_v}(x_v) \right) \cdot \left(\prod_{v \in V \setminus A} \sum_{u_v} d_v(u_v) \delta_{f_v(x_{pa(v) \setminus A}, x'_{pa(v) \cap A}, u_v)}(x_v) \right) \\
 &= \delta_{x'_A}(x_A) \prod_{v \in V \setminus A} k^v(x_{pa(v) \setminus A}, x'_{pa(v) \cap A}; x_v).
 \end{aligned}$$

Summation over all $x_A \in \mathcal{X}_A$ finally gives us

$$\begin{aligned}
 p(x_D \parallel x'_A) &:= \sum_{x_A} p(\widehat{\mathcal{C}}; x_A, x_D) \\
 &= \prod_{v \in D} k^v(x_{pa(v) \setminus A}, x'_{pa(v) \cap A}; x_v).
 \end{aligned}$$

We can rewrite this in the more transparent way by considering one global configuration $x = (x_D, x_A)$ instead of (x_D, x'_A) :

$$p(x_D \parallel x_A) = \prod_{v \in D} k^v(x_{pa(v)}; x_v). \tag{41}$$

This is exactly the truncated product (7).

A.2. Proofs

Proof of Theorem 2. We use the symbols

$$C := \text{an}(A) \quad \text{and} \quad D := V \setminus C.$$

The assumption that B is not a cause of A implies $C \cap B = \emptyset$, and therefore $B \subseteq D$. We prove the theorem in several steps.

Step 1: For the ancestral set C we have

$$p(x_C) = \prod_{v \in C} k^v(x_{pa(v)}; x_v).$$

If $p(x_C) > 0$ we get

$$\begin{aligned}
 p(x_B \parallel x_C) &= \sum_{x_{D \setminus B}} p(x_B, x_{D \setminus B} \parallel x_C) \\
 &= \sum_{x_{D \setminus B}} \prod_{v \in D} k^v(x_{pa(v)}; x_v) \\
 &= \sum_{x_{D \setminus B}} \frac{\prod_{v \in V} k^v(x_{pa(v)}; x_v)}{\prod_{v \in C} k^v(x_{pa(v)}; x_v)} \\
 &= \sum_{x_{D \setminus B}} \frac{p(x_B, x_{D \setminus B}, x_C)}{p(x_C)} \\
 &= \sum_{x_{D \setminus B}} p(x_B, x_{D \setminus B} | x_C) \\
 &= p(x_B | x_C).
 \end{aligned}$$

Step 2: Within this step we are going to prove

$$p(x_B \parallel x_C) = p(x_B \parallel x_A, x_{C \setminus A}) = p(x_B \parallel x_A).$$

In order to do so, we define

$$A' := \{v \in V : \text{there is a path from } C \setminus A \text{ to } v \text{ that does not meet } A\}, \quad B' := V \setminus A'.$$

Clearly we have $C \setminus A \subseteq A'$ and $A \subseteq B'$. Furthermore, the assumption that there is no common cause of A and B also implies $B \subseteq B'$.

$$\begin{aligned}
 p(x_B \parallel x_{C \setminus A}, x_A) &= \sum_{x_{D \setminus B}} p(x_B, x_{D \setminus B} \parallel x_{C \setminus A}, x_A) \\
 &= \sum_{x_{D \setminus B}} \prod_{v \in D} k^v(x_{pa(v)}; x_v) \\
 &= \sum_{x_{A' \setminus C}} \sum_{x_{B' \setminus (A \cup B)}} \prod_{v \in A' \setminus C} k^v(x_{pa(v)}; x_v) \prod_{v \in B' \setminus A} k^v(x_{pa(v)}; x_v) \\
 &= \sum_{x_{B' \setminus (A \cup B)}} \prod_{v \in B' \setminus A} k^v(x_{pa(v)}; x_v) \sum_{x_{A' \setminus C}} \prod_{v \in A' \setminus C} k^v(x_{pa(v)}; x_v) \\
 &= \sum_{x_{B' \setminus (A \cup B)}} \prod_{v \in B' \setminus A} k^v(x_{pa(v)}; x_v) \sum_{x_{A'}} \prod_{v \in A'} k^v(x_{pa(v)}; x_v) \\
 &= \sum_{x_{A'}} \sum_{x_{B' \setminus (A \cup B)}} \prod_{v \in B' \setminus A} k^v(x_{pa(v)}; x_v) \prod_{v \in A'} k^v(x_{pa(v)}; x_v) \\
 &= \sum_{x_{A'}} \sum_{x_{B' \setminus (A \cup B)}} p(x_{A'}, x_{B' \setminus (B \cup A)}, x_B \parallel x_A) \\
 &= p(x_B \parallel x_A).
 \end{aligned}$$

Step 3: Finally,

$$\begin{aligned}
 p(x_B | x_A) &= \sum_{x_{C \setminus A}} p(x_B | x_A, x_{C \setminus A}) p(x_{C \setminus A} | x_A) \\
 &= \sum_{x_{C \setminus A}} p(x_B \parallel x_A, x_{C \setminus A}) p(x_{C \setminus A} | x_A) \quad (\text{Step 1}) \\
 &= \sum_{x_{C \setminus A}} p(x_B \parallel x_A) p(x_{C \setminus A} | x_A) \quad (\text{Step 2}) \\
 &= p(x_B \parallel x_A).
 \end{aligned}$$

This proves the first part of the statement. In the last step we prove the optimality statement.

Step 4: If B is a cause of A or there is a common cause of A and B , then there is a node $v \notin A$ and paths γ, γ' satisfying $v \rightsquigarrow A$ and $v \rightsquigarrow' B$. A Bayesian network $\mathfrak{B}' = (V, E, k')$ with the required properties can be obtained as follows: Consider binary state spaces of the nodes and define local kernels in such a way that all nodes that are not on these paths choose

randomly their state without reference to their parents and with probability $\frac{1}{2}$. The nodes on the paths simply copy the state of v along the edges of the paths. These particular definitions ensure the existence of $x_A, p(x_A) > 0$, and x_B satisfying $p(x_B \parallel x_A) \neq p(x_B|x_A)$. ■

Proof of Theorem 3. We choose a numbering $\{v_1, \dots, v_n\}$ of $\text{an}(A)$ satisfying

$$v_i \in \text{pa}(v_j) \Rightarrow i < j$$

and define $J := \{i \in \{1, \dots, n\} : v_i \in A\}$. With $k := |A|$ we consider the map $\{1, \dots, k\} \rightarrow J, l \mapsto i_l$, that satisfies $1 \leq i_1 < i_2 < \dots < i_k \leq n$. For all $j \in \{1, \dots, k\}$ this implies $\{i_1, \dots, i_{j-1}\} \subseteq \{1, \dots, i_j - 1\}$, and the Markov property gives us the following estimate:

$$\begin{aligned} H(X_A) - \sum_{v \in A} H(X_v | X_{\text{pa}(v)}) &= \sum_{j=1}^k H(X_{v_j} | X_{v_{i_1}}, \dots, X_{v_{i_{j-1}}}) - \sum_{j=1}^k H(X_{v_j} | X_{\text{pa}(v_j)}) \\ &\quad \text{(conditional independence)} \\ &= \sum_{j=1}^k H(X_{v_j} | X_{v_{i_1}}, \dots, X_{v_{i_{j-1}}}) - \sum_{j=1}^k H(X_{v_j} | X_{v_1}, X_{v_2}, \dots, X_{v_{j-1}}) \\ &\geq \mathbb{F}(X_{\text{pa}(v_{i_1})} \rightarrow X_{v_{i_1}}) \\ &\geq \min_{v \in A} \mathbb{F}(X_{\text{pa}(v)} \rightarrow X_v). \end{aligned}$$

This finally implies

$$\begin{aligned} I(X_{v_{i_1}}, \dots, X_{v_{i_k}}) &= \sum_{v \in A} \mathbb{F}(X_{\text{pa}(v)} \rightarrow X_v) - H(X_A) + \sum_{v \in A} H(X_v | X_{\text{pa}(v)}) \\ &\leq \sum_{v \in A} \mathbb{F}(X_{\text{pa}(v)} \rightarrow X_v) - \min_{v \in A} \mathbb{F}(X_{\text{pa}(v)} \rightarrow X_v). \quad \blacksquare \end{aligned}$$

Proof of Eq. (25). We introduce a new symbol

$$\tilde{A}_i := \{v \in V : v \rightsquigarrow v_i \text{ and there exists } j \neq i \text{ satisfying } v_i \not\rightsquigarrow v_j \text{ and } v \rightsquigarrow v_j\}$$

for the right-hand side of (25) and have to prove (see the first definition of (23))

$$\bigcup_{\substack{S \in \mathcal{V} \\ S \supseteq \varphi(v_i)}} \alpha_S = \tilde{A}_i.$$

“ \subseteq ”: We assume that there is an $R \in \mathcal{V}$ with $R \supseteq \varphi(v_i)$ and $v \in \alpha_R$. From $i \in \varphi(v_i)$ we get $i \in R$, and, furthermore, there is a $j \in R \setminus \varphi(v_i)$. This directly implies $v \rightsquigarrow v_i, v_i \not\rightsquigarrow v_j$, and $v \rightsquigarrow v_j$.

“ \supseteq ”: Assume $v \in \tilde{A}_i$. It is sufficient to verify $\varphi(v) \supseteq \varphi(v_i)$: $k \in \varphi(v_i)$ implies $v_i \rightsquigarrow v_k$ and, with $v \rightsquigarrow v_i$ ($v \in \tilde{A}_i$), this implies $v \rightsquigarrow v_k$. Therefore, we have $k \in \varphi(v)$. Now we choose a j with $v_i \not\rightsquigarrow v_j$ and $v \rightsquigarrow v_j$ (such a j exists because $v \in \tilde{A}_i$). This is equivalent to $j \in \varphi(v)$ and $j \notin \varphi(v_i)$. ■

Proof of Theorem 4. Proof of equality chain (29):

It is easy to verify that X_{B_i} is conditionally independent of $X_{A_i \setminus \partial_i^-}$ given $X_{\partial_i^-}$, and that $X_{B_i \setminus \partial_i^+}$ is conditionally independent of X_{A_i} given $X_{\partial_i^+}$ (graph separation criteria for conditional independence, see Section 3.2.2 of [12], or Section 5.3 of [7]):

$$X_{B_i} \perp\!\!\!\perp X_{A_i \setminus \partial_i^-} | X_{\partial_i^-} \quad \text{and} \quad X_{B_i \setminus \partial_i^+} \perp\!\!\!\perp X_{A_i} | X_{\partial_i^+}. \tag{42}$$

This implies

$$\begin{aligned} \mathbb{F}(X_{A_i} \rightarrow X_{B_i}) &= I(X_{A_i} : X_{B_i}) \quad \text{(Examples 2(2))} \\ &= H(X_{B_i}) - H(X_{B_i} | X_{\partial_i^-}, X_{A_i \setminus \partial_i^-}) \\ &= H(X_{B_i}) - H(X_{B_i} | X_{\partial_i^-}) \quad \text{(first conditional independence of (42))} \\ &= H(X_{\partial_i^-}) - H(X_{\partial_i^-} | X_{\partial_i^+}, X_{B_i \setminus \partial_i^+}) \quad \text{(symmetry of mutual information)} \\ &= H(X_{\partial_i^-}) - H(X_{\partial_i^-} | X_{\partial_i^+}) \quad \text{(second conditional independence of (42))} \\ &= I(X_{\partial_i^-} : X_{\partial_i^+}) \\ &= \mathbb{F}(X_{\partial_i^-} \rightarrow X_{\partial_i^+}) \quad \text{(Theorem 2)}. \end{aligned}$$

Proof of the inequalities (30) and (31):

The inequality (30) directly follows from $v_i \in B_i$ for all i . We are going to prove the inequality (31) in several steps. Based on the given Bayesian network, we define a new one by “coarse-graining” and then apply Theorem 3.

Step 1: We consider the set $\mathcal{V} = \{S \subseteq \{1, \dots, n\} : \alpha_S \neq \emptyset\}$ as the node set of a new graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}')$ where two nodes $R, S \in \mathcal{V}$ are connected if there exist $v \in \alpha_R$ and $w \in \alpha_S$ with $(v, w) \in E$, that is

$$\mathcal{E}' := \{(R, S) \in \mathcal{V} \times \mathcal{V} : \text{there is a pair } (v, w) \in E \text{ with } v \in \alpha_R \text{ and } w \in \alpha_S\}.$$

This graph is acyclic because $(R, S) \in \mathcal{E}'$ always implies $S \subsetneq R$: Assume that there exists $i \in S \setminus R$. With $(R, S) \in \mathcal{E}$ there are nodes $v \in \alpha_R, w \in \alpha_S$ with $(v, w) \in E$. According to the definition of the sets α_R and α_S these nodes satisfy $v \not\rightsquigarrow v_i$ and $w \rightsquigarrow v_i$. On the other hand, $(v, w) \in E$ then implies the contradiction $v \rightsquigarrow v_i$.

In order to avoid technicalities, we modify the graph $(\mathcal{V}, \mathcal{E}')$ by adding all pairs $(R, S) \in \mathcal{V} \times \mathcal{V}$ to the edge set \mathcal{E}' if they satisfy $R \supseteq S$. This way, we obtain a graph with extended edge set \mathcal{E} satisfying $\text{pa}(S) = \{R \in \mathcal{V} : R \supseteq S\}$ for all $S \in \mathcal{V}$.

Step 2: For all nodes $S \in \mathcal{V}$ we consider the state set

$$\mathcal{X}_S := \mathcal{X}_{\alpha_S} = \times_{v \in \alpha_S} \mathcal{X}_v$$

by using the original state sets $\mathcal{X}_v, v \in V$. We have the natural identification

$$\mathcal{X}_{\mathcal{V}} \rightarrow \mathcal{X}_{\mathcal{V}}, \quad x = (x_v)_{v \in V} \mapsto \tilde{x} := (x_{\alpha_S})_{S \in \mathcal{V}},$$

and every probability distribution p on $\mathcal{X}_{\mathcal{V}}$ can naturally be considered as probability distribution \tilde{p} on $\mathcal{X}_{\mathcal{V}}$ defined by $\tilde{p}(\tilde{x}) := p(x)$. Given a recursive factorization

$$p(x) = \prod_{v \in V} k^v(x_{\text{pa}(v)}; x_v),$$

we define new kernels k^S as “groups” of the kernels k^v :

$$k^S(\tilde{x}_{\text{pa}(S)}; \tilde{x}_S) := \prod_{v \in \alpha_S} k^v(x_{\text{pa}(v)}; x_v). \tag{43}$$

This obviously provides a recursive factorization of \tilde{p} according to the new graph $(\mathcal{V}, \mathcal{E})$:

$$\tilde{p}(\tilde{x}) = p(x) = \prod_{v \in V} k^v(x_{\text{pa}(v)}; x_v) = \prod_{S \in \mathcal{V}} \left(\prod_{v \in \alpha_S} k^v(x_{\text{pa}(v)}; x_v) \right) = \prod_{S \in \mathcal{V}} k^S(\tilde{x}_{\text{pa}(S)}; \tilde{x}_S).$$

This graph $(\mathcal{V}, \mathcal{E})$, together with the kernels k^S , defines a Bayesian network which we denote by $\tilde{\mathfrak{B}}$.

Step 3: Finally we apply Theorem 3 to the Bayesian network $\tilde{\mathfrak{B}}$ in order to obtain (31):

$$\begin{aligned} I_p(X_{v_1}, \dots, X_{v_n}) &\leq I_p(X_{B_1}, \dots, X_{B_n}) \quad (\text{inequality (30)}) \\ &= I_{\tilde{p}}(X_{\varphi(v_1)}, \dots, X_{\varphi(v_n)}) \quad (\text{Step 2 and } B_i = \alpha_{\varphi(v_i)}) \\ &\leq \max_{j \in \{1, \dots, n\}} \sum_{\substack{i=1 \\ i \neq j}}^n \mathbb{F}_{\tilde{\mathfrak{B}}}(X_{\text{pa}(\varphi(v_i))} \rightarrow X_{\varphi(v_i)}) \quad (\text{Theorem 3}) \\ &= \max_{j \in \{1, \dots, n\}} \sum_{\substack{i=1 \\ i \neq j}}^n \mathbb{F}_{\tilde{\mathfrak{B}}}(X_{A_i} \rightarrow X_{B_i}) \quad (\text{Steps 1 and 2}). \quad \blacksquare \end{aligned}$$

Proof of Corollary 1. The equality chain (32) follows directly from the fact that A_i is an ancestral set (see Examples 2 (2)), and from the conditional independence of X_{v_i} and $X_{A_i \setminus \partial_i^-}$ given $X_{\partial_i^-}$ which is stated in (42). In order to prove the inequality (33) we slightly modify the Bayesian network $\mathfrak{B} = (V, E, k)$ and then apply the corresponding inequality (31).

Step 1: With $A := \bigcup_{i=1}^n A_i$ we define a new graph $\bar{\mathcal{G}} := (\bar{V}, \bar{E})$ by

$$\bar{V} := A \cup \{v_1, \dots, v_n\}, \quad \bar{E} := (E \cap (A \times A)) \cup \bigcup_{i=1}^n (\partial_i^- \times \{v_i\}).$$

Obviously, the parents of a node v_i with respect to this graph $\bar{\mathcal{G}}$ are given by the set ∂_i^- . Denoting the A_i 's and B_i 's that are defined with respect to $\bar{\mathcal{G}}$ by \bar{A}_i and \bar{B}_i we have

$$\bar{A}_i = A_i \quad \text{and} \quad \bar{B}_i = \{v_i\}. \tag{44}$$

Step 2: In order to define the new Bayesian network $\bar{\mathfrak{B}}$, we assign to each node $v_i, i = 1, \dots, n$, a kernel \bar{k}^{v_i} from $\mathcal{X}_{\partial_i^-}$ to \mathcal{X}_{v_i} given by

$$\bar{k}^{v_i}(\mathcal{X}_{\partial_i^-}; \mathcal{X}_{v_i}) := \sum_{x_{B_i \setminus \{v_i\}}} \prod_{v \in B_i} k^v(x_{pa(v)}; \mathcal{X}_v).$$

The other k^v 's where v is an element of A remain unchanged. Obviously, the causal effects of the A_i 's on the v_i 's remain the same, and we have

$$\mathbb{F}_{\bar{\mathfrak{B}}}(X_{A_i} \rightarrow X_{v_i}) = \mathbb{F}_{\mathfrak{B}}(X_{A_i} \rightarrow X_{v_i}). \tag{45}$$

Furthermore, the assumption that $v_i \not\prec v_j$ if $i \neq j$ ensures that the joint probability distribution $p(\bar{\mathfrak{B}})$ coincides with the \bar{V} -marginal of $p(\mathfrak{B})$, that is

$$p(\bar{\mathfrak{B}}; \mathcal{X}_{\bar{V}}) = \sum_{\mathcal{X}_{V \setminus \bar{V}}} p(\mathfrak{B}; \mathcal{X}_{\bar{V}}, \mathcal{X}_{V \setminus \bar{V}}). \tag{46}$$

Step 3: Finally, we prove the inequality (33):

$$\begin{aligned} I_p(X_{v_1}, \dots, X_{v_n}) &= I_{\bar{p}}(X_{v_1}, \dots, X_{v_n}) \quad (\text{Eq. (46)}) \\ &\leq \max_{j \in \{1, \dots, n\}} \sum_{\substack{i=1 \\ i \neq j}}^n \mathbb{F}_{\bar{\mathfrak{B}}}(X_{A_i} \rightarrow X_{B_i}) \quad (\text{inequality (31) of Theorem 4}) \\ &= \max_{j \in \{1, \dots, n\}} \sum_{\substack{i=1 \\ i \neq j}}^n \mathbb{F}_{\mathfrak{B}}(X_{A_i} \rightarrow X_{v_i}). \quad (\text{Eqs. (44) and (45)}). \quad \blacksquare \end{aligned}$$

References

[1] J.J. Atick, Could information theory provide an ecological theory of sensory processing, *Network: Computation in Neural Systems* 3 (2) (1992) 213–251.
 [2] N. Ay, D.C. Krakauer, Geometric robustness theory and biological networks, *Theory in Biosciences* 2 (2007) 93–121.
 [3] N. Ay, D. Polani, Information flows in causal networks, *Advances in Complex Systems* 11 (1) (2008) 17–41.
 [4] H.B. Barlow, Possible principles underlying the transformations of sensory messages, in: W.A. Rosenblith (Ed.), *Sensory Communication: Contributions to the Symposium on Principles of Sensory Communication*, MIT Press, 1959, pp. 217–234.
 [5] H.B. Barlow, Unsupervised learning, *Neural Computation* 1 (1989) 295–311.
 [6] T.M. Cover, J.A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 1991.
 [7] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, D.J. Spiegelhalter, *Probabilistic Networks and Expert Systems*, Springer, 1999.
 [8] D. Edwards, *Introduction to Graphical Modelling*, Springer, 2000.
 [9] A.S. Goldberger, Structural equation models: An overview, in: A.S. Goldberger, O.D. Duncan (Eds.), *Structural Equation Models in the Social Sciences*, Seminar Press, New York, 1973, pp. 1–18.
 [10] A. Klyubin, D. Polani, C. Nehaniv, R.A. Watson, Tracking information flow through the environment: Simple cases of stigmergy, in: *Proceedings of the Ninth International Conference on Artificial Life*, MIT Press, 2004, pp. 563–568.
 [11] S. Laughlin, A simple coding procedure enhances a neuron's information capacity, *Zeitschrift fuer Naturforschung* 36 (1981) 910–912.
 [12] S.L. Lauritzen, *Graphical Models*, Oxford 1996.
 [13] R. Linsker, Self-organization in a perceptual network, *Computer* 21 (3) (1988) 105–117.
 [14] J. Pearl, *Causality: Models, Reasoning and Inference*, Cambridge University Press, 2000.
 [15] H. Reichenbach, *The Direction of Time*, University of California Press, Berkeley, 1956.
 [16] F. Rieke, D. Warland, R. Ruyter van Steveninck, W. Bialek, *Spikes: Exploring the Neural Code*, MIT Press, Cambridge, MA, 1998.
 [17] J. Williamson, *Bayesian Nets and Causality*, Oxford University Press, 2005.
 [18] S. Wright, Correlation and causation, *Journal of Agricultural Research* 20 (1921) 557–585.