Advances in Complex Systems, Vol. 11, No. 1 (2008) 17–41
© World Scientific Publishing Company

INFORMATION FLOWS IN CAUSAL NETWORKS

NIHAT AY
Max Planck Institute for Mathematics in the Sciences,
Inselstrasse 22, D-04103 Leipzig, Germany
and
Santa Fe Institute, 1399 Hyde Park Road,
Santa Fe, New Mexico 87501, USA
nay@mis.mpg.de

DANIEL POLANI
Algorithms and Adaptive Systems Research Groups,
School of Computer Science, University of Hertfordshire,
Hatfield AL10 9AB, UK
d.polani@herts.ac.uk

Received 1 May 2006
Revised 27 June 2007

We use a notion of causal independence based on intervention, which is a fundamental concept of the theory of causal networks, to define a measure for the strength of a causal effect. We call this measure "information flow" and compare it with known information flow measures such as transfer entropy.

Keywords: Causality; information theory; information flow; Bayesian networks.

1. Introduction

> What is mind? No matter.
> What is matter? Never mind.
> — George Berkeley

Information theory provides important quantities for the characterization of complex systems, and there are also some reasons to believe that it pervades the physical world in general [29]. The use of the measure of Shannon's *mutual information* is ubiquitous in this context.

A particular interest lies in the identification of the "flow of information," in the sense of identifying how information is processed in a given system. For this purpose, typically variants of mutual information measures are used [16, 22]. However, much as these measures are used in the context of a "flow of information," they are essentially of correlative character. This, in particular, creates some situations where such quantities are difficult to be interpreted as a "flow." The utility of having a proper measure for a "flow of information" can be seen in a number of recent papers that use simplified forms of information flow measures to characterize complexity of information processing [4,27], robustness [3], or information processing in agents [10,11], as well as the complexity of neural interactions [24]. Thus, the variety of applications for a notion of information flow signals an increased need for a well-founded measure of information flow and promises a wide and fruitful scope of applications for such a measure.

How to go about constructing such a measure? As we mentioned above, a pure correlative measure does not precisely fit the bill. Different parts of a system may share information (i.e. have mutual information), but without information flowing between these parts. Rather, the joint information stems from a common past.

For an intuitive picture of how to move toward a measure of information flow, consider for example a river whose waterflow one wishes to track. The standard method of tracking the waterflow is to introduce a tracer (color or radioactivity) into the river and to trace the occurrence of this tracer throughout the river [28]. Central to the success of the method is that the tracer consists of a material with distinctive properties not usually found in the river.

In a similar mode, one could try to trace down information in a system. Given an information processing system, one would add ("inject," [15]) some noise uncorrelated with any of the unperturbed parts of the system and measure the mutual information of different parts of the perturbed system with the noise [24]. Since the noise is uncorrelated with the unperturbed system (corresponding to the tracer material not found in the river before the measurement), any mutual information found is an indicator for an information flow. This "active probing" [18] is at the core of the experimental method. To imbue such an intuition with a precise meaning, it is necessary to have a formal notion of causality [19]. This was already noted by Lloyd [15] who recognized the importance of causal structures and interventions for constructing a meaningful notion of information flow.

Note that there is a central difference between measuring the flow of information and the flow of matter (as in the river example). Matter flows are additive. This allows one to estimate the unperturbed flows via infinitesimal perturbations of the system. Information flows, however, are nonadditive. Thus, one cannot expect naive "active probing" to be a suitable direct measure for the information flow in an unperturbed system. This task of calculating the information flow will occupy us for the rest of this paper.

As with the models of material flow, we will employ graph models. The realization of the information-theoretic perspective is achieved by considering the nodes of this graph to be random variables. An appropriate formalism, (*causal*) *Bayesian networks*, is well developed. The above "injection" of information is modeled in this context as *intervention* in a given network, i.e. as a modification of the original network [19]. In particular, this is intimately connected with a thoroughly studied framework for the treatment of causal dependencies. The concept of information flow that we will develop on the basis of causal Bayesian networks can be seen as an information-theoretic counterpart of the probabilistic formalism from Ref. 19.

As in Ref. 19, we will consider Bayesian networks with a finite number of nodes each of which assumes a finite discrete number of states. While it is difficult to say whether the formalism generalizes easily to systems with infinitely many nodes (such as a continuous set of nodes), we expect the formalism to extend naturally to the case where the state spaces of the nodes themselves may be continuous.[a]

2. Directed Acyclic Graphs

We consider a finite and nonempty set V of *nodes* and a set $E \subseteq V \times V$ of *edges*[b] between the nodes. Such a *directed graph* $G := (V, E)$ serves as a model for the causal interactions of the nodes, and we write $v \to w$ if $(v, w) \in E$. Two nodes v, w are *adjacent*, in symbols $v \sim w$, if $v \to w$ or $w \to v$. We write $\mathsf{pa}(v) := \{w \in V : (w, v) \in E\}$ for the set of parents of v. An ordered sequence $(v_1, \ldots, v_k)$ of nodes is called a *path* from v_1 to v_k if $v_i \sim v_{i+1}$ for all $i = 1, \ldots, k-1$. A path is *directed* if it satisfies $v_i \to v_{i+1}$ for all $i = 1, \ldots, k-1$. If $v_1 = v_k$, the directed path is called *directed cycle*. A directed graph without directed cycles is called a *directed acyclic graph* (*DAG*).

In his graphical models approach to causality, Pearl [19] assumes a DAG as the structural specification of a causal network. Within this approach one aims at understanding the relation between these structural and the corresponding observational properties, such as stochastic dependence or independence of the nodes. In this regard, *d*-separation (*d* stands for "*directional*") has been identified as the graphical separation property that is consistent with stochastic conditional independence (see Theorem 1). It is defined as follows:

Definition 1 (*d*-separation). Let $G = (V, E)$ be a DAG, and let S be a subset of V. We say that a path $(v_1, \ldots, v_k)$ is *blocked* by S if there is a node v_i on the path such that:

- either $v_i \in S$, and edges of the path do not meet head-to-head at v_i, or
- v_i and all its descendants are not in S, and edges of the path meet head to head at v_i.

A set A is *d-separated* from B by S if all paths from A to B are blocked by S. Throughout the paper, d-separation will refer to disjoint sets A, B, S.

This somewhat complicated notion of separation turns out to be the optimal graphical test for the conditional stochastic independence of two variables A and B given a third variable S. This optimality property has been proven by Verma and Pearl ([19]; see Theorem 1 in Sec. 3). In order to have a better intuitive understanding of the d-separation concept, we consider a simple example where S is the empty set. In that case, a set A is d-separated from a set B (by the empty set) if every

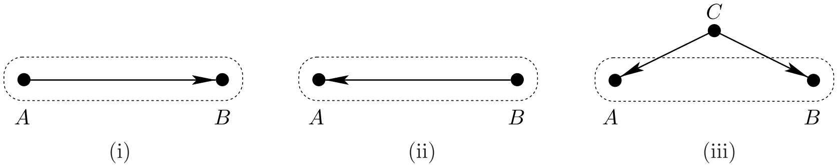
[a] An information transfer formalism for continuous spaces that attempts to implement some causal aspects was introduced in Ref. 13.

[b] Note that since the edges are denoted by 2-tuples, they imply directedness.

path from A to B has a head-to-head node. Stated differently, A and B are not d-separated if we have one of the following three situations:

(i) A is a cause of B, or
(ii) B is a cause of A, or
(iii) there is a common cause C of A and B, i.e. a set $C \subseteq V \backslash (A \cup B)$, and directed paths γ and γ' from C to A and C to B.

These three situations are known to provide the graphical basis for stochastic dependence of A and B.



While the notion of d-separation is characterized by its consistency with stochastic conditional independence structures, the causal interpretation of arrows as direct causal relations suggests another separation concept. Intuitively, one would like to call a variable B causally independent of A, if A is not a cause of B, or, stated differently, B is not an effect of A. In the graphical representation, this means that there is no directed path from A to B. The graphical criterion for causal conditional independence would then be given by a *unidirectional* separation condition as the graphical representation of causal conditional independence structures, which we call *ud*-separation (this definition is equivalent to the path interception concept by Galles and Pearl [8], but more convenient for our exposition):

Definition 2 (*ud*-separation). Let $G = (V, E)$ be a DAG, and let A, B, S be three disjoint subsets of V. We say that B is *ud-separated* from A by S (in G) if all directed paths from A to B go through S.

Example 1 (DAG layers). Let $G = (V, E)$ be a DAG. We stratify the set V in a natural way into layers. We start with $V_1 := \{v \in V : \mathsf{pa}(v) = \emptyset\}$. Obviously, V_1 is not empty, because otherwise we could construct a directed cycle. In order to get the next layers we iterate according to

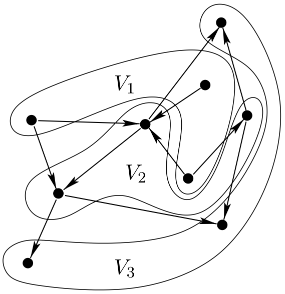
$$V_{k+1} := \{v \in V \backslash (V_1 \cup \ldots \cup V_k) : \mathsf{pa}(v) \cap (V_1 \cup \ldots \cup V_k) \neq \emptyset\}, \qquad k = 1, 2, \ldots.$$

For some k, V_{k+1} is an empty set, and therefore all sets $V_{k+2}, V_{k+3}, \ldots$ are also empty. With $L := \max\{k : V_k \neq \emptyset\}$ we have the disjoint union

$$V = V_1 \cup \ldots \cup V_L$$

and the corresponding map $l : V \to \{1, \ldots, L\}$ which assigns to each $v \in V$ its layer number $l(v)$.

Note that this definition does not imply a strict feed-forward architecture. By definition of the layers we know that for $l(v_{i+1}) > l(v_i)$ we always have $l(v_{i+1}) = l(v_i)+1$, i.e. forward edges always increase the layer index by exactly 1. However, there is no such constraint for backward edges and edges inside a single layer.



Now, it turns out that for $1 \leq r < s < t \leq L$, the layer V_t is *ud*-separated from V_r by V_s. In order to see this, consider a directed path $(v_1, \ldots, v_k)$ from V_r to V_t. Then the corresponding layer numbers $l(v_1), l(v_2), \ldots, l(v_k)$ start with r and end with t. This implies that the numbers have to go through s, and therefore the path $(v_1, \ldots, v_k)$ meets V_s.

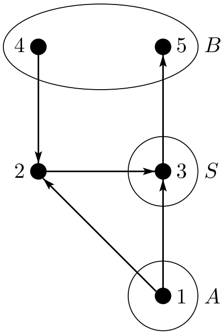
Proposition 1. *Let $G = (V, E)$ be a DAG, and let A, B, S be three disjoint subsets of V. If B is d-separated from A by S, then B is also ud-separated from A by S.*

Proof. Let $(v_1, \ldots, v_k)$ be a directed path from A to B. The *d*-separation property implies that this path is blocked by S. Because all nodes in the path are head-to-tail, i.e. $\to v_i \to$, the only way for the path to be blocked by S is that there exists a $v_i \in S$. □

Example 2. Consider the set $V := \{1, 2, 3, 4, 5\}$ of nodes and the set

$$E := \{(1,2), (1,3), (2,3), (4,2), (3,5)\}$$

of edges, as shown in the following figure:



Furthermore, $A := \{1\}, B := \{4, 5\}, S := \{3\}$. Obviously, B is *ud*-separated from A by S but not *d*-separated, because the path $(1, 3, 2, 4)$ is not blocked by S.

3. Causal Models

In Sec. 2 we presented the structural model for causal interactions. In order to quantitatively specify these interactions, we need a description of the nodes' information processing. We assume that each node $v \in V$ has a nonempty and finite set $\mathcal{X}_v$ of states. Given a subset A, the *configurations in* A are the elements of the set $\mathcal{X}_A := \prod_{v \in A} \mathcal{X}_v$, and one has the natural projection $X_A : \mathcal{X}_V \to \mathcal{X}_A$, $x = (x_v)_{v \in V} \mapsto x_A := (x_v)_{v \in A}$. We now describe the nodes v by Markov kernels

$$p_v : \mathcal{X}_{\mathsf{pa}(v)} \times \mathcal{X}_v \to [0, 1], \qquad (x_{\mathsf{pa}(v)}, x_v) \mapsto p_v(x_v | x_{\mathsf{pa}(v)}).$$

Within Pearl's causality theory, these Markov kernels are interpreted mechanistically as a generative model for the local processing of the nodes. In other words, the kernels do not just reflect an observed probabilistic relation between the values of a node and its parents, but an actual (mechanistic) dependence: if the values of the parent nodes are manipulated (i.e. intervention takes place), this affects their child node directly.

This interpretation can be additionally motivated by the so-called *structural equation modeling* technique. In this classical statistical modeling method, one considers deterministic functions together with hidden random disturbances. It turns out that this can be easily related to the above kernel formalism. In particular, the causal aspects do not depend on a particular representation by structural equations. Therefore, given a DAG G, a family of local kernels p_v, $v \in V$, is called a *G-causal model.* From such a mechanistic model one can then move to the phenomenological level, obtaining the probability of observing a global configuration $x = (x_v)_{v \in V}$ if all nodes v are generating their output according to the kernels p_v. This joint distribution is then given by

$$p(x) = \prod_{v \in V} p_v(x_v | x_{\mathsf{pa}(v)}). \tag{1}$$

We have the following central theorem by Verma and Pearl [19, 26].

Theorem 1 (Verma and Pearl). *Let $G = (V, E)$ be a DAG, and let A, B, S be three disjoint subsets of V. Then B is d-separated from A by S if and only if for all G-causal models X_A and X_B are stochastically independent given X_S* [*with respect to the joint distribution* (1)].

This theorem establishes the connection between the underlying graphical structure of a causal model and the corresponding stochastic independence structure with respect to the joint distribution. The deviation from stochastic independence can be quantified by information-theoretic measures like mutual information, conditional mutual information, or multi-information, a generalization of mutual information to several variables [17]. This way, the qualitative nature of stochastic independence is embedded in a quantitative theory which provides a measure identifying stochastic interdependencies among the nodes. Unfortunately, in applications this is often confused with the identification of causal relationships. In this paper we present a quantitative theory of causal dependence that is based on the notion of *ud* separation instead of *d*-separation. Theorem 2, an alternative formulation of Theorem 11 from Ref. 8, will be an analog to Theorem 1. In what follows we need the notion of causal effects [19], which is based on the possibility of intervention in causal models. For didactical reasons we define causal effects in two steps:

Step 1. Basically, we split the node set V into a subset C of nodes in which we intervene and into the subset D of remaining nodes which are observed. Let x_C be some (fixed) configuration in C. Setting $X_C = x_C$ means replacing all mechanisms p_v, $v \in C$, in (1) by constant values x_v, $v \in C$ deriving from the configuration x_C. A transparent representation of the corresponding post-interventional distribution is obtained by considering the probability of observing a configuration x_D in the complement $D := V \setminus C$ of C after having set x_C.

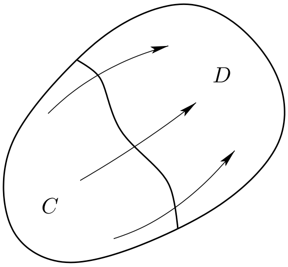
$$p(x_D|\hat{x}_C) \;:=\; \prod_{v \in D} p_v(x_v|x_{\mathsf{pa}(v)}). \tag{2}$$

By putting a "hat" on x_C we want to distinguish the notation of this kind of conditioning from the standard conditioning in probability theory, which for $p(x_C) > 0$ is given by

$$p(x_D|x_C) \;=\; \frac{p(x_D, x_C)}{p(x_C)}. \tag{3}$$

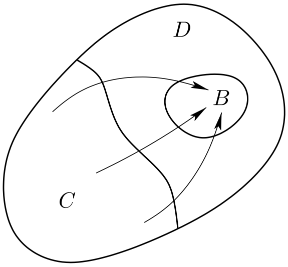
In order to verbally distinguish these two types of conditioning, we will refer to them as *interventional* and *observational* conditioning.

Compared with the pre-interventional distribution (1), the post-interventional distribution (2) is obtained just by neglecting all factors p_v, where v is an element of C (*truncated factorization*). Note that this interventional conditioning, in contrast to observational conditioning, is defined for *all* $x_C \in \mathcal{X}_C$. The map $(x_C, x_D) \mapsto p(x_D|\hat{x}_C)$ is called *direct causal effect* $C \to D$, as indicated in the following figure:



For a subset A of C and a (fixed) configuration $x_{C\setminus A} \in \mathcal{X}_{C\setminus A}$, we call the map $(x_A, x_D) \mapsto p(x_D|\hat{x}_A, \hat{x}_{C\setminus A})$ *direct causal effect* $A \to D$ *imposing* $x_{C\setminus A}$.

Step 2. In order to deal with causal effects that are mediated by some uncontrolled variables, we consider an arbitrary subset B of D as shown here:



The probability of observing $X_B = x_B$ after having set $X_C = x_C$ by intervention is given by

$$p(x_B|\hat{x}_C) \;=\; \sum_{x_{D\setminus B}} p(x_B, x_{D\setminus B}|\hat{x}_C) \;=\; \sum_{x_{D\setminus B}} \prod_{v\in D} p_v(x_v|x_{\mathsf{pa}(v)}). \tag{4}$$

The corresponding map $(x_C, x_B) \mapsto p(x_B|\hat{x}_C)$ is called *causal effect* $C \to B$. As with the direct effects of the first step, we consider a subset A of C and a configuration $x_{C\setminus A} \in \mathcal{X}_{C\setminus A}$. The map $(x_A, x_B) \mapsto p(x_B|\hat{x}_A, \hat{x}_{C\setminus A})$ is the *causal effect* $A \to B$ *imposing* $x_{C\setminus A}$. This allows us to compute the probability distribution of x_B depending on the particular intervention x_A if the variables $x_{C\setminus A}$ are fixed beforehand (by intervention, not observation).

4. Causal Independence

We want to study causal independence. To this end, let us first have a look at the usual concept of stochastic independence. Let A, B, S be three disjoint subsets of V. Then X_A and X_B are stochastically independent given X_S if for all x_A, x_S with positive probability $p(x_A, x_S)$ and all x_B

$$p(x_B|x_A, x_S) = \sum_{x'_A} p(x'_A|x_S)p(x_B|x'_A, x_S) \quad [= p(x_B|x_S)]. \tag{5}$$

This condition means that observing x_A after having observed x_S does not change our expectation of observing x_B. In analogy, we formulate an interventional version of this: Setting x_A after having set x_S does not change the probability of observing x_B. This corresponds to the condition:

$$p(x_B|\hat{x}_A, \hat{x}_S) = \sum_{x'_A} p(x'_A|\hat{x}_S)p(x_B|\hat{x}'_A, \hat{x}_S). \tag{6}$$

Unlike the conditional probability $p(x_B|x_A, x_S)$, the interventional probability $p(x_B|\hat{x}_A, \hat{x}_S)$ is defined for *all* pairs x_S, x_A rather than being limited to those with positive probability. This is due to the fact that interventional probabilities are defined via mechanisms rather than observations. Being able to formulate this stronger condition allows us to define that X_B *is causally independent of* X_A *imposing* X_S, written as

$$X_B \perp\!\!\!\perp X_A | \widehat{X}_S$$

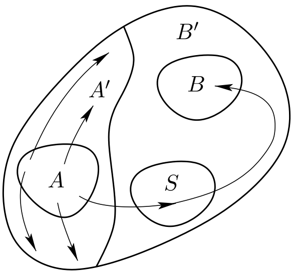
if condition (6) is fulfilled for all pairs x_A, x_S. Note that this specifically includes situations of "unseen" or "unprobed" causal dependence; this is always defined because the network mechanisms are given for the whole domain of their input variables. Furthermore, note that the causal independence property is not symmetric. This is consistent with our intuitive understanding of causality as a directional concept. In particular, this notion of independence is governed by rules that are different from those underlying a graphoid structure [8]. Graphoids are an abstract mathematical structure that captures the rules of graph-induced stochastic independence, particularly symmetry, in compact form.

Now we are ready for Theorem 2, a causal analog to Theorem 1. It will relate the *ud*-separation property associated with the graphical structure of a causal model to causal independence. The latter depends on the specification of the local (mechanistic) conditional probabilities. We present an alternative formulation and proof to Theorem 11 from Ref. 8.

Theorem 2. *Let $G = (V, E)$ be a DAG, and let A, B, S be disjoint subsets of V. Then B is ud separated from A by S if and only if for all G-causal models X_B is causally independent of X_A imposing X_S.*

Proof. "Only if": We assume that B is *ud*-separated from A by S, and set $D := V \backslash (A \cup S)$. We are going to prove that $p(x_B|\hat{x}_A, \hat{x}_S)$ does not depend on x_A. To this end we define

$$A' := \{v \in V : \text{there exists a directed path from } A \text{ to } v \text{ that does not meet } S\},$$
$$B' := V \backslash A'.$$



By definition one has $A \subseteq A'$ and $S \subseteq B'$. Furthermore, the assumption of *ud*-separation implies $B \subseteq B'$. Thus, we can decompose D into a disjoint union of the sets $A' \backslash A$ and $B' \backslash S$. Now we are ready to prove that $p(x_B|\hat{x}_S, \hat{x}_A)$ does not depend on x_A:

$$\begin{aligned}
p(x_B|\hat{x}_A,\hat{x}_S) &= \sum_{x_{D\setminus B}} p(x_B, x_{D\setminus B}|\hat{x}_A,\hat{x}_S) \\
&= \sum_{x_{D\setminus B}} \prod_{v\in D} p_v(x_v|x_{\mathsf{pa}(v)}) \\
&= \sum_{x_{A'\setminus A}} \sum_{x_{B'\setminus(S\cup B)}} \prod_{v\in A'\setminus A} p_v(x_v|x_{\mathsf{pa}(v)}) \prod_{v\in B'\setminus S} p_v(x_v|x_{\mathsf{pa}(v)}) \\
&= \sum_{x_{B'\setminus(S\cup B)}} \prod_{v\in B'\setminus S} p_v(x_v|x_{\mathsf{pa}(v)}) \underbrace{\sum_{x_{A'\setminus A}} \prod_{v\in A'\setminus A} p_v(x_v|x_{\mathsf{pa}(v)})}_{=1} \\
&= \sum_{x_{B'\setminus(S\cup B)}} \prod_{v\in B'\setminus S} p_v(x_v|x_{\mathsf{pa}(v)}).
\end{aligned}$$

The definition of A' and B' implies that for all $v \in B'\setminus S$ one has $\mathsf{pa}(v) \subset B'$. Therefore all the expressions $p_v(x_v|x_{\mathsf{pa}(v)})$ of the last line, and therefore also $p(x_B|\hat{x}_A,\hat{x}_S)$, do not depend on x_A, which implies Eq. (6).

"If": We assume that X_B is causally independent of X_A imposing X_S for all G-causal models and want to prove that B is *ud*-separated from A by S. We define $\mathcal{X}_v := \{0,1\}$ for all $v \in V$. Assume that there is a directed path $(v_1,\ldots,v_k)$ from A to B. Without loss of generality we can assume that $v_i \notin A \cup B$ for all $1 < i < k$. Every node v_i, $i = 2,\ldots,k$, just copies the state of v_{i-1}, which is contained in the set $\mathsf{pa}(v_i)$:

$$p_{v_i}(x_{v_i}|x_{\mathsf{pa}(v_i)}) := \delta_{x_{v_{i-1}}}(x_{v_i}) := \begin{cases} 1 & \text{if } x_{v_i} = x_{v_{i-1}}, \\ 0 & \text{otherwise.} \end{cases}$$

All other nodes are assumed to choose their state completely randomly according to $p_v(x_v|x_{\mathsf{pa}(v)}) := \frac{1}{2}$.

$$\begin{aligned}
p(x_B|\hat{x}_S,\hat{x}_A) &= \sum_{x_{D\setminus B}} \prod_{v\in D} p_v(x_v|x_{\mathsf{pa}(v)}) \\
&= \sum_{x_{D\setminus B}} \prod_{i=2}^{k} p_{v_i}(x_{v_i}|x_{\mathsf{pa}(v_i)}) \prod_{v\in D\setminus\{v_2,\ldots,v_k\}} p_v(x_v|x_{\mathsf{pa}(v)}) \\
&= \frac{1}{2^{|D|-k+1}} \sum_{x_{D\setminus B}} \prod_{i=2}^{k} p_{v_i}(x_{v_i}|x_{\mathsf{pa}(v_i)}) \\
&= \frac{1}{2^{|D|-k+1}} \sum_{x_{D\setminus B}} \delta_{x_{v_1}}(x_{v_2})\delta_{x_{v_2}}(x_{v_3})\cdots\delta_{x_{v_{k-1}}}(x_{v_k}) \\
&= \frac{1}{2^{|B|-1}}\, \delta_{x_{v_1}}(x_{v_k}).
\end{aligned}$$

Thus $p(x_B|\hat{x}_S, \hat{x}_A)$ clearly depends on x_A, and therefore X_B is not causally independent of X_A imposing X_S. □

Combined with Theorem 1 this result directly implies the following corollary:

Corollary 1. *Let G be a DAG, and let A, B, S be three disjoint subsets of V. If for all G-causal models X_B is stochastically independent of X_A given X_S, then for all G-causal models X_B is causally independent of X_A imposing X_S.*

Proof. Stochastic independence for all G-models is, according to Theorem 1, equivalent to *d*-separation. On the other hand, according to Proposition 1, *d*-separation implies *ud*-separation and therefore causal independence. □

5. A Definition of Information Flow

We now proceed to quantify causal dependence. For this, we first have to look at the stochastic dependence case. Stochastic dependence is measured by deviation from independence — more precisely, the deviation of the left hand side of (5) from its right hand side. For this purpose, we need to specify a measure of deviation or distance between transition kernels. The application of the *relative entropy* as such a measure turns out to be very consistent with information-theoretic concepts. With a probability distribution p on $\mathcal{X}_C$, the relative entropy of two transition kernels P and Q from $\mathcal{X}_C$ to $\mathcal{X}_B$ is defined as

$$D_p(P\|Q) := \sum_{x_C} p(x_C) \sum_{x_B} P(x_B|x_C) \log \frac{P(x_B|x_C)}{Q(x_B|x_C)}.$$

Here we apply the usual convention that $0 \log \frac{0}{r} = 0$ and $s \log \frac{s}{0} = \infty$ for all $r \geq 0$ and all $s > 0$. Throughout the paper, log stands for the binary logarithm $\log_2$. Using this deviation measure, the stochastic dependence of X_A and X_B given x_S is quantified as the deviation from independence [i.e. condition (5)].

$$I_p(X_A : X_B|x_S) := \sum_{x_A} p(x_A|x_S) \sum_{x_B} p(x_B|x_A, x_S) \log \frac{p(x_B|x_A, x_S)}{\sum_{x'_A} p(x'_A|x_S) p(x_B|x'_A, x_S)}. \tag{7}$$

Taking the mean with respect to $p(x_S)$, $x_S \in \mathcal{X}_S$, gives us

$$I_p(X_A : X_B|X_S) = \sum_{x_S} p(x_S) I_p(X_A : X_B|x_S). \tag{8}$$

This is called the *conditional mutual information* of X_A and X_B given X_S. In the case where S is the empty set, this quantity reduces to the *mutual information* $I_p(X_A : X_B)$. One has the property

$$X_B \perp\!\!\!\perp X_A | X_S \Leftrightarrow I_p(X_A : X_B|X_S) = 0.$$

Now let us come back to causal dependence. Similarly to (7), we define it as deviation from causal independence, which is given by Eq. (6): The causal contribution of X_A to X_B imposing x_S is measured by

$$I_p(X_A \to X_B | \hat{x}_S) := \sum_{x_A} p(x_A|\hat{x}_S) \sum_{x_B} p(x_B|\hat{x}_A, \hat{x}_S) \log \frac{p(x_B|\hat{x}_A, \hat{x}_S)}{\sum_{x'_A} p(x'_A|\hat{x}_S) p(x_B|\hat{x}'_A, \hat{x}_S)}.$$

By taking the mean, we finally obtain the *information flow from X_A to X_B imposing X_S*:

$$I_p(X_A \to X_B | \widehat{X}_S) := \sum_{x_S} p(x_S) I_p(X_A \to X_B | \hat{x}_S).$$

It has the same structure as (8), and it is a measure for the “visible” contribution of a causal effect. In the case where S is empty we simply write $I_p(X_A \to X_B)$, in analogy with the mutual information. It should be noted that the information flow measure can be reformulated in terms of conditional mutual information with respect to a modified distribution:

$$\hat{p}(x_S, x_A, x_B) := p(x_S) p(x_A|\hat{x}_S) p(x_B|\hat{x}_S, \hat{x}_A). \qquad (9)$$

With this definition we have

$$I_p(X_A \to X_B | \widehat{X}_S) = I_{\hat{p}}(X_B : X_A | X_S).$$

Proposition 2.

$$X_B \perp\!\!\!\perp X_A | \widehat{X}_S \;\Rightarrow\; I_p(X_A \to X_B | \widehat{X}_S) = 0. \qquad (10)$$

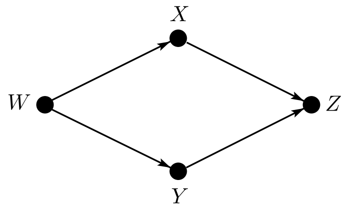
If $\hat{p}$ as defined in (9) *is strictly positive, then the implication* (10) *becomes an equivalence.*

Proof. This follows directly from the well-known properties of the relative entropy. □

A combination of this statement with Theorem 2 directly implies the following:

Corollary 2. *If $I_p(X_A \to X_B | \widehat{X}_S) > 0$, then there exists a directed path from A to B that does not meet S.*

Example 3 (diamond structure). Consider the following graph with nodes $V = \{W, X, Y, Z\}$ and edges $E = \{(W, X), (W, Y), (Y, Z), (X, Z)\}$:



We assume that all nodes have as state set $\{0, 1\}$. Node W generates a state w with probability $p_1(w) = \frac{1}{2}$, which is then copied by nodes X and Y. Finally, node Z generates the XOR value of the two states x and y, which, in this case, is always 0. These mechanisms give us the joint distribution

$$p(w, x, y, z) = \frac{1}{2}\delta_w(x)\delta_w(y)\delta_{\mathrm{XOR}(x,y)}(z).$$

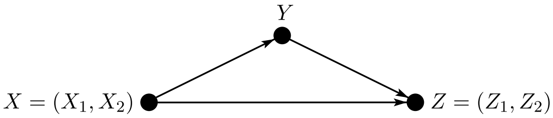
By straightforward calculations we obtain the following quantities, which illustrate that, in general, our measures of correlation and causation express different aspects of the system:

Correlation	Causation
$I_p(X : Y) = 1$	$I_p(X \to Y) = 0$
$I_p(X : Y \mid W) = 0$	$I_p(X \to Y \mid \widehat{W}) = 0$
$I_p(W : Z \mid Y) = 0$	$I_p(W \to Z \mid \widehat{Y}) = 1$

The example shows that in moving from the correlational to the causal measure, the value can go up as well as down.

Example 4 (channel splitting). Consider three nodes $X = (X_1, X_2)$, Y, and $Z = (Z_1, Z_2)$. Node X generates a pair $(x_1, x_2) \in \{0, 1\} \times \{0, 1\}$ with probability $p_X(x_1, x_2)$. One entry — say, x_1 — is copied by Z_1. The second entry x_2 first goes to Y and then to Z_2. This gives the joint distribution

$$p(x_1, x_2, y, z_1, z_2) = p_X(x_1, x_2)\delta_{x_2}(y)\delta_{x_1}(z_1)\delta_y(z_2).$$



An easy calculation shows that the information flow from X to Z imposing Y coincides with the entropy $H_p(X_1)$ of X_1:

$$I_p(X \to Z|\widehat{Y}) = H_p(X_1).$$

If Y were not imposed, then the total flow from X to Z would just be $H_p(X)$, i.e. the full entropy of the input node X.

Example 5 (mediated flow). Consider the graph shown in Example 3 with nodes W, X, Y, and Z. Again, W generates a symbol $w \in \{0, 1\}$ with probability $\frac{1}{2}$, which is then copied by nodes X and Y. For node Z we consider two cases:

Case 1. Z is assumed to copy the state from X, and we have the joint distribution

$$p(w, x, y, z) = \frac{1}{2}\delta_w(x)\delta_w(y)\delta_x(z). \tag{11}$$

The conditional mutual information $I_p(X : Z|Y)$ vanishes, because X and Y provide the same information for Z. On the other hand, our information flow measure $I_p(X \to Z|\widehat{Y})$ has the maximum achievable value of 1 bit. Note that this is equal to the unintervened information flow $I_p(X \to Z)$.

Case 2. We modify the mechanism of Z for the counterfactual situation where X and Y are different. In that situation Z is now assumed to generate a symbol $z \in \{0, 1\}$ with probability $\frac{1}{2}$. The mechanism for identical x and y remains as in the first case. We have the joint distribution

$$p(w, x, y, z) = \frac{1}{2}\delta_w(x)\delta_w(y) \cdot \begin{cases} \delta_x(z) & \text{if } x = y, \\ \dfrac{1}{2} & \text{if } x \neq y, \end{cases} \tag{12}$$

which coincides with the joint distribution (11) of the first case. But, here, Y determines to some extent whether X can control the outcome of Z. More precisely, one has

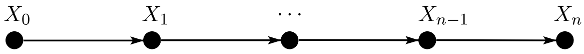
$$I_p(X \to Z|\widehat{Y}) = \frac{3}{4} \log \frac{4}{3} \approx 0.31.$$

The result lies significantly below the maximum achievable information flow of 1 bit due to the mediating effect of the imposed variable Y.

6. Information Flows in Markov Chains

Consider a chain $X_0, X_1, X_2, \dots, X_n$ that is generated by an intitial distribution p_0 and a (fixed) transition kernel p_X. In this case we have the joint distribution

$$p(x_0, x_1, \dots, x_n) = p_0(x_0)p_X(x_1|x_0)p_X(x_2|x_1) \cdots p_X(x_n|x_{n-1}).$$



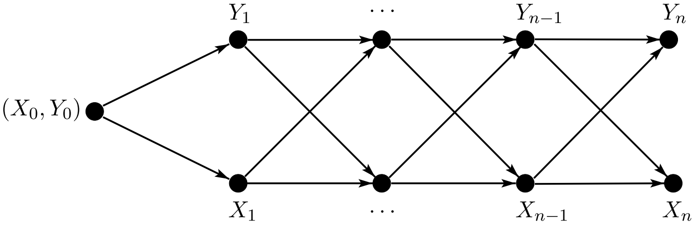
There is a field of research [16] which is not restricted to this simple setting, but also deals with more general dynamical systems and aims at relating the qualitative characteristics of a given dynamics to its information flow in time. Hereby, information flow is usually quantified by the mutual information between a time interval $[i, j] = \{i, i + 1, \dots, j\}$ in the past and a time interval $[k, l] = \{k, k + 1, \dots, l\}$ in the future. Applied to our simple example, this would correspond to the mutual information

$$I_p(X_{[i,j]} : X_{[k,l]}), \quad 1 \leq i \leq j < k \leq l \leq n. \tag{13}$$

Within the context of the present paper, it is natural to ask whether our definition of information flow is consistent with the definition (13). Indeed, a small calculation proves that

$$I_p(X_{[i,j]} \to X_{[k,l]}) = I_p(X_{[i,j]} : X_{[k,l]}).$$

However, this consistency breaks down if one wants to quantify information flows among the elements of a composite dynamical system, i.e. a dynamical system partitioned into different subsystems. To make clear in what sense this is meant, we consider a system consisting of two interacting processes X and Y, as shown in the following figure:



The processes are assumed to be generated by an initial joint distribution p_0, separated into variables X_1 and Y_1 and then propagated via kernels p_X and p_Y, as follows:

$$\begin{aligned} & p(x_0, \cdots, x_n, y_0, \cdots, y_n) \\ & \quad = p_0(x_0, y_0)\delta_{x_0}(x_1)\delta_{y_0}(y_1)p_X(x_2|x_1, y_1)p_Y(y_2|x_1, y_1) \cdots \\ & \qquad p_X(x_n|x_{n-1}, y_{n-1})p_Y(y_n|x_{n-1}, y_{n-1}). \end{aligned}$$

Schreiber [22] has proposed a measure, called *transfer entropy*, that, applied to this situation, is intended to be capable of quantifying the information transfer from Y to X. For $1 \leq i, j \leq k < n$, it is defined as the conditional mutual information $I_p(Y_{[i,k]} : X_{k+1}|X_{[j,k]})$. The following simple but instructive example ($i = j = k$) proves that the transfer entropy does not necessarily coincide with the information flow $I_p(Y_k \to X_{k+1}|\widehat{X}_k)$:

Example 6 (information exchange). We consider two observationally equivalent cases:

Case 1. Assume that the nodes X_i, Y_i in both node sequences can each assume the state -1 or $+1$, and assume that at each time step k they just copy the state of the node in the other sequence, i.e. $p(x_{k+1}|x_k, y_k) = \delta_{y_k}(x_{k+1})$ and vice versa.

We start with a configuration (x_0, y_0), distributed according to the probability distribution $\frac{1}{2}(\delta_{(-1,+1)} + \delta_{(+1,-1)})$. For each such initial configuration, we will, over time, observe an (X, Y) sequence of the form $\cdots \to (-1, +1) \to (+1, -1) \to (-1, +1) \to \cdots$. The transfer entropy vanishes in this case for all times k (because the predictability of the next state of component X_{k+1} from its earlier state X_k is perfect). This contradicts the intuition that by copying the state from the other node sequence, there clearly *is* a flow of information. On the other hand, our measure of information flow has indeed the maximal value of 1 bit in this case, consistent with intuition.

Case 2. Consider now the case where, after initial generation as in Case 1, X_{k+1} is the inversion of X_k for all k (i.e. -1 becomes $+1$, and $+1$ becomes -1) and, likewise, Y_{k+1} is the inversion of Y_k. In particular, there is no interaction between sequences X and Y after their initial generation. This is observationally entirely equivalent to Case 1 and thus the transfer entropy remains 0, as before. However, its interventional dynamics is fundamentally different, and the information flow $I_p(Y_k \to X_{k+1}|\widehat{X}_k)$ becomes 0 in this case. Thus information flow is able to distinguish between the case of information being actively exchanged between the chains X and Y and the case where there is no such exchange, as intuitively expected.

Example 7 (soft transmission). We consider the following two extreme situations:

$$p(x_k|x_{k-1}, y_{k-1}) = p(y_k|x_{k-1}, y_{k-1}) = \frac{1}{2}. \tag{14}$$

Clearly, the information flow from the current state of Y to the next state of X imposing the current state of X here is zero. The other extreme situation is given in the following way: in order to compute the next state, both nodes copy the current state of node Y and invert it.

$$(x, y) \to (-y, -y), \quad x, y \in \{\pm 1\}. \tag{15}$$

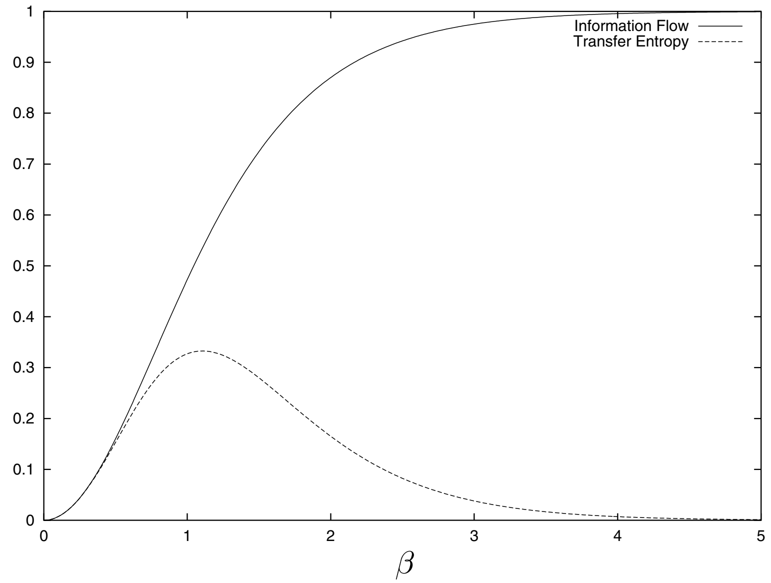
In this case, the current state of Y completely determines the next state of X. Therefore, intuitively, one would expect a maximal amount of information flow, as shown in the following diagram:

$$\begin{tikzcd}
Y_0 \arrow[r] \arrow[dr] & Y_1 \arrow[r] \arrow[dr] & \cdots \arrow[r] \arrow[dr] & Y_{k-2} \arrow[r] \arrow[dr] & Y_{k-1} \arrow[r] \arrow[dr] & Y_k \\
X_0 & X_1 & \cdots & X_{k-2} & X_{k-1} & X_k
\end{tikzcd}$$

We now interpolate these two extreme situations of minimal and maximal information flow in order to get a one-parameter family of transition kernels. More precisely, we define

$$p(y_k|x_{k-1}, y_{k-1}) := \frac{1}{1 + e^{2\beta y_k y_{k-1}}}, \qquad p(x_k|x_{k-1}, y_{k-1}) := \frac{1}{1 + e^{2\beta x_k y_{k-1}}}.$$

Here β plays the role of an inverse temperature. In the high-temperature limit ($\beta \to 0$) we recover the completely random transition (14), and in the low-temperature limit ($\beta \to \infty$) we recover the transition (15). The following diagram compares the shape of the information flow $I_p(Y_{k-1} \to X_k|\widehat{X}_{k-1})$ and the transfer entropy $I_p(Y_{k-1} : X_k|X_{k-1})$ with respect to the unique stationary distribution p as a function of the inverse temperature β.



As we can see, the information flow is consistent with the intuition that moving from $\beta = 0$ to $\beta = \infty$ corresponds to an interpolation between a transition with vanishing information flow and a transition with maximal information flow. Near $\beta = 0$ the transfer entropy increases as β becomes larger and is close to the information flow. But for larger β's it starts to decrease again and converges to zero for $\beta \to \infty$. The reason for that is simply that the transition for large β generates more redundancy between the two processes X and Y. Therefore, as β grows, an increasing amount of information about Y_{k-1} can be computed from information about X_{k-1}, which lets the transfer entropy $I_p(X_k : Y_{k-1} | X_{k-1})$ decrease toward zero. More precisely, we have the following Markov transition matrix describing the global dynamics [the rows denote (x_{k-1}, y_{k-1}), the columns (x_k, y_k) and the entries the transition probabilities from a state at time $k-1$ to time k]:

	$(-1,-1)$	$(+1,-1)$	$(-1,+1)$	$(+1,+1)$
$(-1,-1)$	a^2	ab	ab	b^2
$(+1,-1)$	a^2	ab	ab	b^2
$(-1,+1)$	b^2	ab	ab	a^2
$(+1,+1)$	b^2	ab	ab	a^2

where

$$a := \frac{1}{1+e^{2\beta}}, \quad b := \frac{1}{1+e^{-2\beta}}.$$

As stationary distribution we get

$$p(+1,+1) = p(-1,-1) = \frac{1}{2} - ab, \quad p(-1,+1) = p(+1,-1) = ab.$$

As we can see, for $\beta = 0$ we have the uniform distribution $p(x, y) = \frac{1}{4}$, which implies that there is no correlation between the two nodes. As β increases, we get more and more redundancy, and in the limit $\beta \to \infty$ we get totally correlated nodes with distribution $\frac{1}{2}(\delta_{(-1,-1)} + \delta_{(+1,+1)})$.

7. Intervention "Without Intervention"

Upto now we used the term "intervention" in a purely mathematical sense, without reference to its intrinsically operational meaning. Furthermore, we assumed that the causal model, as a mathematical structure, is given and therefore can be used for the calculation of causal effects and corresponding information flows. On the other hand, within an experimental setting the causal model is in general not known, and experimental intervention in the real system provides the direct way to identify causal effects. The effects of this active intervention would then allow the calculation of causal information flows. The transfer entropy, in contrast, although not an adequate measure for information flow, is determined by purely observational quantities and does not require any intervention. This raises the question whether one also can derive the effect of an active intervention by observational data only, i.e. without active intervention. Since there are observationally equivalent joint distributions induced by different causal models, it is clear that the solution of this identifiability problem in general requires some additional knowledge about the system. The specification of this additional knowledge represents a central point in Pearl's theory.

In order to be more precise, we consider two disjoint subsets A and B and intend to compute the causal effect $p(x_B|\hat{x}_A)$. Given a (strictly positive) joint distribution $p(x_B, x_A)$ one can compute the conditional distribution $p(x_B|x_A)$. This is in general not sufficient for the computation of the causal effect $p(x_B|\hat{x}_A)$. Now, instead of $p(x_A, x_B)$, we consider the joint distribution $p(x_A, x_B, x_C)$ on an extended system $S := A \cup B \cup C$, and ask the question whether this is sufficient for the computation of $p(x_B|\hat{x}_A)$. If this is the case, the causal effect is called *identifiable* in S. In particular, the complete extension $C := V \backslash (A \cup B)$ is always sufficient. On the other hand, we want to find a minimal extension. We mention one special case in order to illustrate this point. Consider two nodes v and w and set $A = \{v\}$, $B = \{w\}$, and $C := \mathsf{pa}(v)$. Given a strictly positive distribution $p(x_{\mathsf{pa}(v)}, x_v, x_w)$, the so-called *adjustment for direct causes* provides an expression of the causal effect $p(x_w|\hat{x}_v)$:

$$p(x_w|\hat{x}_v) = \sum_{x_{\mathsf{pa}(v)}} p(x_w|x_v, x_{\mathsf{pa}(v)})p(x_{\mathsf{pa}(v)}). \tag{16}$$

The expressions on the right hand side of (16) require only the knowledge of the marginal $p(x_{\mathsf{pa}(v)}, x_v, x_w)$. Formula (16) allows one to calculate the causal effect of node v on node w if only *all* parents of v are known. Note that (16) is in general different from the usual conditional

$$p(x_w|x_v) = \sum_{x_{\mathsf{pa}(v)}} p(x_w|x_v, x_{\mathsf{pa}(v)})p(x_{\mathsf{pa}(v)}|x_v).$$

In the following, three simple scenarios of simulated systems demonstrate the calculation of the information flow based on observational data only and also provide further examples of how it differs from the transfer entropy.

Scenario 1. Consider the following diagram. Let X_0 be a random variable that assumes the values -1 and $+1$ with probability $p(X_0 = x) = \frac{1}{2}$ for each $x \in \{-1, +1\}$, and analogously for Y_0.[c]

$$\begin{tikzcd}
X_0 \arrow[r] \arrow[dr] & X_1 \arrow[r] \arrow[dr] & \cdots \arrow[r] \arrow[dr] & X_{k-2} \arrow[r] \arrow[dr] & X_{k-1} \arrow[r] \arrow[dr] & X_k \\
Y_0 & Y_1 & \cdots & Y_{k-2} & Y_{k-1} & Y_k
\end{tikzcd}$$

Now consider both the horizontal and the diagonal arrows to indicate a *copy* process, i.e.

$$\begin{aligned}
p(x_k|x_{k-1}) &= \delta_{x_{k-1}}(x_k), \\
p(y_k|x_{k-1}) &= \delta_{x_{k-1}}(y_k).
\end{aligned}$$

In order to ensure strict positivity, we replace this exact copy process by the following perturbed copy process:

$$p(x_k|x_{k-1}) = (1-\varepsilon)\delta_{x_{k-1}}(x_k) + \varepsilon\frac{1}{2}, \tag{17}$$

$$p(y_k|x_{k-1}) = (1-\varepsilon)\delta_{x_{k-1}}(y_k) + \varepsilon\frac{1}{2}, \tag{18}$$

where $\varepsilon = 5 \times 10^{-3}$. We now desire to calculate the information flow from X_{k-1} to Y_k imposing Y_{k-1} (for $k \geq 2$) from experimental observation and compare it with the corresponding conditional mutual information (information transfer). In our experiment, we generate samples for X_0 and Y_0 and apply the perturbed copy dynamics. From the concrete realizations of the corresponding stochastic process we estimate (via relative frequencies) the conditional distribution $p(y_k|x_{k-1}, y_{k-1})$ and the joint distribution $p(x_{k-1}, y_{k-1})$. This information allows for an approximative estimate of the conditional mutual information $I_p(Y_k : X_{k-1}|Y_{k-1})$. In order to estimate the causal information flow, we have to consider the conditional mutual information with respect to the joint distribution

$$\hat{p}(y_{k-1}, x_{k-1}, y_k) := p(y_{k-1})p(x_{k-1}|\hat{y}_{k-1})p(y_k|\hat{x}_{k-1}, \hat{y}_{k-1}). \tag{19}$$

[c] Note that, as the diagram indicates, X_0 and Y_0 are independent.

Without knowing whether X is influencing Y or the other way round, we can compute the interventional quantities on the right hand side of (19) by observational information with the help of (16):

$$p(x_{k-1}|\hat{y}_{k-1}) = p(x_{k-1}), \quad p(y_k|\hat{x}_{k-1}, \hat{y}_{k-1}) = p(y_k|x_{k-1}, y_{k-1}).$$

In our concrete application we generate 10^6 samples and apply our estimation for $k = 3$. As a result, the information flow $I_p(X_{k-1} \to Y_k|\widehat{Y}_{k-1})$ is about 0.97 bits (it is not exactly 1 because of the noise introduced to ensure full support) while the transfer entropy $I_p(X_{k-1} : Y_k|Y_{k-1})$ is 0.039 bits (again not exactly 0 because of the small noise). This shows that the directedness of the diagram is reflected immediately in the information flow. The transfer entropy fails to see the flow from the X to the Y row because it essentially is a measure of predictability. Since the values of Y at time steps earlier than Y_k are confounded with Y_k because of the peculiar copying process, the transfer entropy does not see much new information in Y_k that cannot be predicted by Y_{k-1}, and thus virtually vanishes. On the other hand, the information flow is only preoccupied with the actual causal structure of the system, and identifies the flow from X_{k-1} to Y_k as ≈ 1 bit, in agreement with intuition.

Scenario 2. In the second experiment, the setup is almost the same as in the previous, except that we use different arrows. Again, X_0 and Y_0 are (independently) equally distributed over -1 and $+1$, and the arrows again indicate copying (with the same slight amount of noise as before).

$$\begin{tikzcd}
X_0 \arrow[r] & X_1 \arrow[r] & \cdots \arrow[r] & X_{k-2} \arrow[r] & X_{k-1} \arrow[r] & X_k \\
Y_0 & Y_1 & \cdots & Y_{k-2} & Y_{k-1} & Y_k
\end{tikzcd}$$

As expected, both transfer entropy $I_p(X_{k-1} : Y_k|Y_{k-1})$ and information flow $I_p(X_{k-1} \to Y_k|\widehat{Y}_{k-1})$ were close to 0 (order of magnitude of 2×10^{-6} bits, due to sampling noise of the experiment).

Scenario 3. In the last experiment of the series, we calculate the same flows as before. However, the directionality of the graph is exactly the opposite to that in the first experiment: the arrows point from Y to X.

$$\begin{tikzcd}
X_0 & X_1 & \cdots & X_{k-2} & X_{k-1} & X_k \\
Y_0 \arrow[r] \arrow[ur] & Y_1 \arrow[r] \arrow[ur] & \cdots \arrow[r] \arrow[ur] & Y_{k-2} \arrow[r] \arrow[ur] & Y_{k-1} \arrow[r] \arrow[ur] & Y_k
\end{tikzcd}$$

Everything else is kept the same, including the noisy copying operation. The transfer entropy, still based on prediction, is again close to 0 (1×10^{-6} bits). But, unlike in the first scenario, the information flow is also now close to 0 (around 3×10^{-6} bits), indicating the nonexisting causal connection between X_{k-1} and Y_k.

As a side remark, the mutual information $I_p(X_{k-1} : Y_k)$ is ≈ 0.97 bit in the first scenario, ≈ 0 bit in the second and ≈ 0.94 bit in the third. It consistently overestimates both transfer entropy (which consists in just correcting for the overestimate of mutual information due to the correlated past) and information flow. We thus see that the information flow captures our intuitive picture of an example where other values do not give us the intuitively expected directionality.

Again, we emphasize that all information flow quantities in Scenarios 1–3 were calculated from purely observational data, without having to *carry out* any actual intervention. Note, however, that even while we use only observational data, the resulting information flow measure is still a fundamentally interventional quantity. For this to be possible, we have made use of the causal structure of the system; in this case, we used the fact that the nodes X_{k-2}, Y_{k-2} formed the complete parent set of X_{k-1}, Y_{k-1}.

8. Further Application Areas

In Sec. 1, we have given a brief outline for possible applications of the concept of information flow. The sketched scenarios reflect problems with a concrete and immediate need for an information flow quantity that measures causal information transmission in a system. Furthermore, recent interest in measures such as Schreiber's transfer entropy or Granger causality (which, despite its name, is an essentially correlative quantity) demonstrates that the issue of quantifying causal aspects of information diffusion is of high timeliness and relevance for current research. The information flow measure introduced in the present paper addresses this question utilizing Pearl's full-fledged causal framework. In this section, we wish to give a flavor of possible applications for information flow.

Physics. The unambiguous causal interpretation of information flow allows one to enhance the identification of causal relations and mechanisms in general physical systems. Measuring their impact provides a new tool for quantitative studies of dynamical and complex systems, in which information-theoretic quantities have already played a significant role in the past.

Synchronization. Synchronization is a phenomenon of great interest in the context of self-organization [23]. Our expectation is that the information flow formalism can help elicit which information flows between the different components of a system are involved and necessary for achieving the effects of global synchronization. This could help to clarify the mechanisms that allow the synchronization of movement or behavior of, for example, distributed biological systems.

Game dynamics. Often one encounters game-theoretic scenarios with a dynamic component, i.e. two players that adapt their strategies over time or two populations where the distribution of available strategies changes during evolution [21]. Of particular relevance are dynamics moving toward cooperative or antagonistic player behavior. In creating models for, say, social or economic dynamics, it would be of significant interest to use information flow to attribute how much a given player is "responsible" for the emergence of a particular cooperative or antagonistic outcome. This could provide a deeper understanding of some of the drives governing socioeconomic systems.

Models for the perception–action loop. In Sec. 1, some work using information-flow-type quantities has been mentioned. Information-theoretic principles, long believed to be relevant for the understanding of biological information processing [2, 6] are beginning to receive renewed attention [5, 14]. Related to that, Bayesian and prediction-based concepts of the self-organization of the perception–action loop prove to be increasingly successful [7, 12, 20]. The family of information flow methods thus promises to provide a calculus by which some principles guiding the emergence and development of biological (and artificial) perception–action loops of intelligent agents can be identified and formulated [9, 10], providing a generalization of the information-theoretic treatment of control systems [1, 25].

With its causal character, the concept of information flow provides an additional tool in this arsenal of methods and could help to elucidate further important issues relevant to the information processing dynamics in biological and artificial agents.

9. Conclusions

This work was motivated by the need for a systematic quantification of the "flow of information." In developing this concept, we desired to capture, on the one hand, essential properties of a Shannon-type quantity measurable in bits. On the other hand, we aimed to realize a flow-like philosophy different from the correlative nature of the notion of mutual information.

This required us to deviate from the computation of mutual information which is based on purely observational quantities. An adequate modification of the formalism to reflect the "flow" aspect required us to take into account the causal nature of the systems under study. For this, we used the interventional formalism from Ref. 19, which provided an appropriate framework for treating the causal mechanisms in a given system. First, we formulated the classical mutual information by quantifying the deviation of two random variables from stochastic independence. Then, information flow was introduced, in analogy with mutual information, as the deviation of two random variables from causal independence. For this purpose we had to appropriately modify the probabilistic quantities involved in establishing stochastic independence to fit a causal framework, with the help of Pearl's interventional calculus.

In a number of examples we have shown that our measure for information flow is indeed different from other widely used notions, such as transfer entropy or other quantities related to mutual information; in particular, our information flow is indeed able to distinguish cases in an intuitive way which observational methods cannot distinguish (Example 6), and has demonstrated that transfer entropy and information flow can differ in a nontrivial way (Example 7).

Together with information flow, we have presented an appropriate modification of well-established formalisms and concepts fitting the framework of causal Bayesian networks. This allowed us to demonstrate how the notion of information flow is embedded in a broad and robust framework of conceptual tools.

The concept of causality and information flow shows nicely how drastically the ability to intervene (or “experiment”) modifies our understanding of the world. Particularly striking is the fact that, while observational quantities are easier to obtain (no experiments are needed), the causal concept of *ud*-separation is more intuitive than the observational concept of *d*-separation; this expresses Pearl's philosophy that causal knowledge seems to be significantly less brittle than observational (probabilistic) knowledge [19].

New notions are typically introduced as generalizations or adaptations of existing and established concepts, often driven by theoretical considerations. However, one of the strongest justifications for introducing a new notion is the actual practical need for a notion with suitable properties. This was exactly the case for information flow. If well constructed, such a notion cannot just help to cover the cases that motivated its introduction, but also open up pathways toward novel insights into systems not previously considered. The conceptual framework and the scenarios studied in the present paper indicate that information flow may be a promising candidate for achieving this.

Acknowledgments

The authors thank A. Bell, N. Bertschinger, J. Jost, D. Krakauer, E. Olbrich, Y. Sato, F. Sommer, T. Wennekers, A. Klyubin and C. Nehaniv for the many fruitful discussions on the subject of information flow. Nihat Ay thanks the Santa Fe Institute for supporting him as an external faculty member.

References

[1] Ashby, W. R., *Design for a Brain* (Wiley & Sons, New York, 1952).

[2] Atick, J. J., Could information theory provide an ecological theory of sensory processing?, *Netw. Comp. Neural Syst.* **3** (1992) 213–251.

[3] Ay, N. and Krakauer, D. C., Information geometric theories for robust biological networks, *Theor. Biosci.* **125** (2007) 93–121.

[4] Ay, N. and Wennekers, T., Dynamical properties of strongly interacting Markov chains, *Neural Netw.* **16**(10) (2003) 1483–1497.

[5] Baddeley, R., Hancock, P. and Földiák, P. (eds.), *Information Theory and the Brain* (Cambridge University Press, 2000).

[6] Barlow, H. B., Possible principles underlying the transformations of sensory messages, in *Sensory Communication: Contributions to the Symposium on Principles of Sensory Communication*, ed. Rosenblith, W. A. (M.I.T. Press, 1959), pp. 217–234

[7] Der, R., Steinmetz, U. and Pasemann, F., Homeokinesis: A new principle to back up evolution with learning, in *Computational Intelligence for Modelling, Control, and Automation*, ed. Mohammadian, M., Concurrent Systems Engineering Series, Vol. 55 (IOS Press, 1999), pp. 43–47.

[8] Galles, D. and Pearl, J., Axioms of causal relevance, *Artif. Int.* **97** (1997) 9–43.

[9] Klyubin, A. S., Polani, D. and Nehaniv, C., Representations of space and time in the maximization of information flow in the perception-action loop, *Neural Comp.* **19** (2007) 2387–2432.

[10] Klyubin, A. S., Polani, D. and Nehaniv, C. L., Organization of the information flow in the perception-action loop of evolved agents, in *Proc. 2004 NASA/DoD Conference on Evolvable Hardware* (IEEE Computer Society, 2004), pp. 177–180.

[11] Klyubin, A. S., Polani, D. and Nehaniv, C. L., Empowerment: A universal agent-centric measure of control, in *Proc. IEEE Congress on Evolutionary Computation*, 2–5 Sep. 2005, Edinburgh, Scotland (CEC '05) (IEEE, 2005), pp. 128–135.

[12] Körding, K. P. and Wolpert, D. M., Bayesian integration in sensorimotor learning, *Nature* **427** (2004) 244–247.

[13] Liang, X. S. and Kleeman, R., Information transfer between dynamical system components, *Phys. Rev. Lett.* **95** (2005) 244101.

[14] Linsker, R., Self-organization in a perceptual network, *Computer* **21** (1988) 105–117.

[15] Lloyd, S., Causality and information flows, in *Information Dynamics*, eds. Atmanspacher, H. and Scheingraber, H. (Plenum Press, 1991), pp. 131–142.

[16] Matsumoto, K. and Tsuda, I., Calculation of information flow rate from mutual information, *J. Phys. A: Math. Gen.* **21** (1988) 1405–1414.

[17] McGill, W., Multivariate information transmission, *IEEE Trans. Infor. Theor.* **4** (1954) 93–111.

[18] Olsson, L., Nehaniv, C. L. and Polani, D., Sensory channel grouping and structure from uninterpreted sensor data, in *IEEE NASA/DoD Conference on Evolvable Hardware 2004* (IEEE Computer Society, 2004), pp. 153–160.

[19] Pearl, J., *Causality: Models, Reasoning and Inference* (Cambridge University Press, 2000).

[20] Porr, B., Egertony, A. and Wörrgötter, F., Towards closed loop information: Predictive information, *Constr. Foundations* **1**(2) (2006).

[21] Sato, Y. and Ay, N., Adaptive dynamics of interacting Markovian processes. Working Paper 06-12-051 (Santa Fe Institute, 2006).

[22] Schreiber, T., Measuring information transfer, *Phys. Rev. Lett.* **85** (2000) 461–464.

[23] Strogatz, S., *Sync: The Emerging Science of Spontaneous Order* (Theia, 2004).

[24] Tononi, G. and Sporns, O., Measuring information integration, *BMC Neurosci.* **4** (2003) 31.

[25] Touchette, H. and Lloyd, S., Information-theoretic approach to the study of control systems, *Physica A* **331** (2003) 140–172.

[26] Verma, T. and Pearl, J., Causal networks: Semantics and expressiveness, in *Proc. Fourth Workshop on Uncertainty in Artificial Intelligence* (Mountain View, California, USA, 1988), pp. 352–359.

[27] Wennekers, T. and Ay, N., Finite state automata resulting from temporal information maximization, *Neural Comp.* **17** (2005) 2258–2290.

[28] Werner, A., Hötzl, H., Käss, W. and Maloszewski, P., Interpretations of tracer experiments in the Danube–Aach–system, Western Swabian Alb, Germany, with

analytical models, in *Karst Waters & Environmental Impacts*, eds. Günay and Johnson (Balkema, Rotterdam, 1997), pp. 153–160.

[29] Wheeler, J. A., Information, physics, quantum: The search for links, in *Complexity, Entropy and the Physics of Information*, ed. Zurek, W. H., Santa Fe Studies in the Sciences of Complexity (Addison-Wesley, Reading, Massachusetts, 1990), pp. 3–28.