INFORMATION FLOWS IN CAUSAL NETWORKS

NIHAT AY[1] AND DANIEL POLANI[2]

Abstract. We introduce a notion of causal independence based on virtual intervention, which is a fundamental concept of the theory of causal networks. Causal independence allows for defining a measure for the strength of a causal effect. We call this information flow and compare it with known information flow measures such as the transfer entropy.

Contents

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 (Wheeler, 1990). 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 as to identify how information is processed in a given system. For this purpose, typically variants of mutual information measures are used (Shaw, 1981, 1984; Matsumoto and Tsuda, 1988; Schreiber, 2000). However, as 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 quntities 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 (Wennekers and Ay, 2005), robustness (Ay and Krakauer, 2006), or information

[1]N. Ay: Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany & Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

[2]D. Polani: Algorithms and Adaptive Systems Research Groups, School of Computer Science, University of Hertfordshire, Hatfield AL10 9AB, United Kingdom

Date: May 12, 2006.

processing in agents (Klyubin et al., 2004, 2005). 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 how to move towards a measure of information flow, consider e.g. a river whose waterflow one wishes to track. The standard method to track the waterflow is to introduce a tracer (color or radioactivity) into the river and to trace the occurrence of this tracer throughout the river (Werner et al., 1997). Central for the success of the method is that the tracer consists of a material 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", Klyubin et al., 2006) 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. 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.

There is, however, a central difference to measuring the flow of matter (as in the river illustration). Matter flows are additive. This allows to estimate the unperturbed flows via infinitesimal perturbations of the system. Information flows, however, are non-additive. Thus, one can not expect naive "active probing" to be a suitable direct measure for the information flow in an unperturbed system (Klyubin et al., 2006). This task of calculating the information flow in the unperturbed system will occupy us for the rest of this paper.

Similar to 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. The formalism to do so, *(causal) Bayesian networks*, is well developed. Above "injection" of information is modeled in this context as *intervention* in a given network, i.e. as a modification of the original network (Pearl, 2000). In particular, this is intimately connected with a thoroughly studied framework for the treatment of causal dependencies (Lauritzen, 2005, 1996). 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 (Pearl, 2000).

As in (Pearl, 2000), we will consider Bayesian networks with a finite number of nodes who take on a finite discrete number of states. While it is difficult to say whether the formalism generalizes easily to systems with continuous nodesets, we expect the formalism to generalize naturally to the case where the state spaces for the nodes may be continuous.

2. Directed Acyclic Graphs

We consider a finite set $V \neq \emptyset$ of *nodes* and a set $E \subseteq V \times V$ of *edges* 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 $w \to v$. An ordered sequence $(v_1, \ldots, v_k)$ is called a *path* from 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 (Pearl, 2000) assumes DAG as the structural specification of causal networks. 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: We say that a path $(v_1, \ldots, v_k)$ is *blocked* by a set S, if there is a node v_i of 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. While this condition is characterized by its consistency with stochastic conditional independence structures, Pearl's notion of *causality* suggests an *unidirectional* separation condition as graphical representation of causal conditional independence structures, which we call *ud*-separation:

Definition 1 (*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. If this is the case, we write $(B \perp_{ud} A \,|\, S)_G$ or, to simplify notation, $B \perp_{ud} A \,|\, S$.

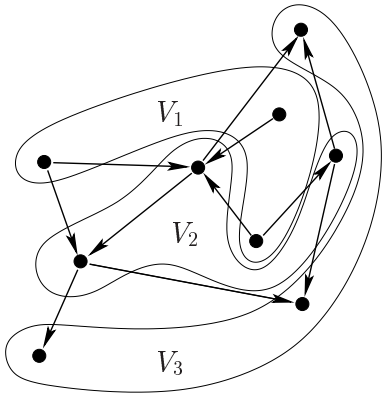
Example 1 (DAG Layers). Let $G = (V, E)$ be a DAG. We stratify the set V in a natural way into layers. We start with $V_1 := \{v \in V : \mathrm{pa}(v) = \emptyset\}$. Obviously, V_1 is not empty, because otherwise we could construct a directed cycle. In order to get the next layers we iterate according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

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

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

$$V = V_1 \cup \cdots \cup V_L$$

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



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. By definition of the layers we know that for $l(v_{i+1}) > l(v_i)$ we always have $l(v_{i+1}) = l(v_i) + 1$. This implies that the numbers have to go through s, and therefore the path $(v_1, \ldots, v_k)$ meets V_s.

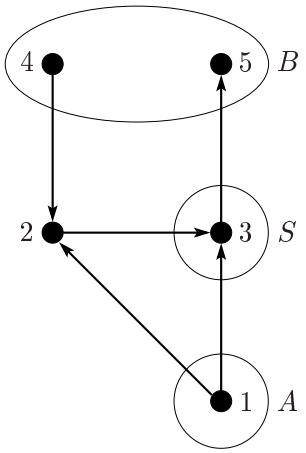
Proposition 1. *Let $G = (V, E)$ be a DAG, and let A, B, S be three disjoint subsets of V. If B is d-separated from A by S, then B is also ud-separated from A by S.*

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, that is $\to v_i \to$, 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.

3. Causal Models

In Section 2 we presented the structural model for causal interactions. In order to specify these ineractions we need a concrete mechanistic description of the nodes. We assume that each node $v \in V$ has a non-empty 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 canonical projections $X_A : \mathcal{X}_V \to \mathcal{X}_A$, $x = (x_v)_{v \in V} \mapsto x_A := (x_v)_{v \in A}$. We now describe the mechanisms of the nodes v by Markov kernels

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

Given a DAG G, we call a family of local kernels p_v, $v \in V$, a *G-causal model*. The corresponding joint distribution is then given by

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

We have the following central theorem by Verma and Pearl (Pearl, 2000):

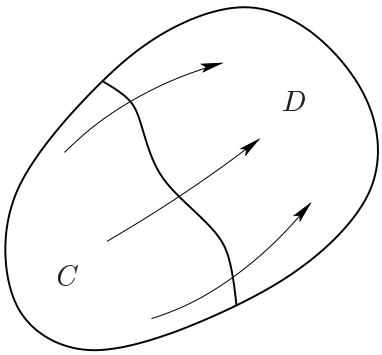
Theorem 1 (Verma & Pearl, 1988). *Let $G = (V, E)$ be a DAG, and let A, B, S be three disjoint subsets of V. Then B is d-separated from A by S if and only if for all G-causal models X_A and X_B are stochastically independent given X_S (with respect to the joint distribution (1)).*

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. This way, the qualitative nature of stochastic independence is embedded in a quantitative theory, which allows for the identification of stochastic interdependencies among the nodes. In applications this is often misinterpreted as identification of causal relationships. In this paper we present a quantitative theory of causal dependence that is based on our notion of *ud*-separation instead of *d*-separation. Theorem 2, our main result, will be an analogon to Theorem 1. In what follows we need the notion of causal effects (Pearl, 2000), which is based on the possibility to intervene 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 that are intervened and the subset D of remaining nodes which are observed. Let x_C be a configuration in C. Setting $X_C = x_C$ means replacing all mechanisms p_v, $v \in C$, in (1) by the constants x_v, $v \in C$. 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.

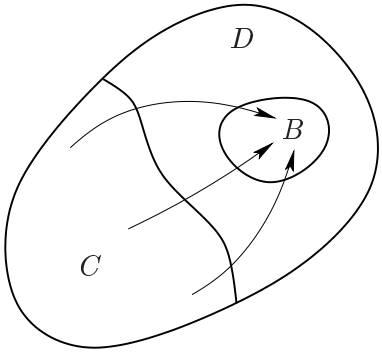
$$p(x_D|\hat{x}_C) := \prod_{v \in D} p_v(x_v|x_{\mathrm{pa}(v)}) \tag{2}$$

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 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

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

The corresponding map $(x_C, x_B) \mapsto p(x_B|\hat{x}_C)$ is called *causal effect* $C \to B$. Similar to 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* $x_{C\setminus A}$.

4. Causal Independence

We want to study causal independence. To this end, first let us have look at 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

$$p(x_B|x_A, x_S) \;=\; \sum_{x'_A} p(x'_A|x_S)\, p(x_B|x'_A, x_S) \quad \Big(\,=\; p(x_B|x_S)\Big) \tag{3}$$

This condition means that observing x_A after having observed x_S does not change our expectation of observing x_B. An interventional version of this would be: Setting x_A after having set x_S does not change the probability of observing x_B. This corresponds to the following condition:

$$p(x_B|\hat{x}_A, \hat{x}_S) \;=\; \sum_{x'_A} p(x'_A|\hat{x}_S)\, p(x_B|\hat{x}'_A, \hat{x}_S) \tag{4}$$

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

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

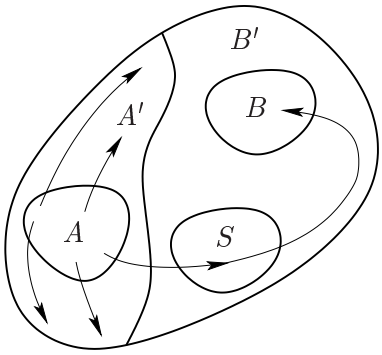
if condition (4) is fulfilled for all pairs x_S, x_A. Note that this specifically includes situations of "unseen" or "unprobed" causal dependence, which is induced by the network mechanisms. Furthermore, note that the causal independence property is not symmetric. This is consistent with our intuitive 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 (Pearl, 2000).

Now we are ready for our main result of the paper, which, in analogy to Theorem 1, relates the *ud*-separation property associated with the graphical structure of a causal model to the causal independence relation, which depends on the specification of the local conditional probabilities.

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 $D := V \setminus (A \cup S)$. 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

$$A' := \{v \in V \ : \text{ there exists a directed path from } A \text{ to } v \text{ that doesn't meet } S\}, \quad B' := V \setminus A'.$$



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

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

The definition of A' and B' implies that for all $v \in B' \setminus S$ one has $\mathrm{pa}(v) \subset B'$. Therefore all the expressions $p_v(x_v|x_{\mathrm{pa}(v)})$ of the last line, and therefore also $p(x_B|\hat{x}_A, \hat{x}_S)$, do not depend on x_A, which implies equation (4).
"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 not intersecting S. Without restriction of generality we can assume $v_i \notin A \cup B$ 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 $\mathrm{pa}(v_i)$:

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

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

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

Thus $p(x_B|\hat{x}_A, \hat{x}_S)$ 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-causal models is, according to Pearl, 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

In order to quantify causal dependence we first have look at the stochastic dependence case. Stochastic dependence is measured by deviation from independence, more precisely, the deviation of the left-hand side of (3) from its right-hand side. In order to do so, 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.

$$I_p(X_A : X_B \,|\, x_S) \;:=\; \sum_{x_A} p(x_A|x_S) \sum_{x_B} p(x_B|x_A, x_S) \log \frac{p(x_B|x_A, x_S)}{\sum_{x'_A} p(x'_A|x_S)\, p(x_B|x'_A, x_S)} \tag{5}$$

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

$$I_p(X_A : X_B \,|\, X_S) \;=\; \sum_{x_S} p(x_S)\, I_p(X_A : X_B \,|\, x_S) \tag{6}$$

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 following property

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

Now let us come back to causal dependence. Similarily to (5) we define it as deviation from causal independence, which is given by equation (4): 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 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 (6), and it is a measure for the "visible" contribution of a causal effect. In the extreme case where S is empty the information flow quantifies the total causal effect which is mediated by all variables in $V \setminus (A \cup B)$, and we simply write $I_p(X_A \to X_B)$ in analogy to the mutual information. In the other extreme case where S is the complement of A and B in V the information flow quantifies the direct causal effect $A \to B$.

Proposition 3.

$$X_B \perp\!\!\!\perp X_A \,|\, \widehat{X}_S \quad \Rightarrow \quad I_p(X_A \to X_B \,|\, \widehat{X}_S) \;=\; 0 \tag{7}$$

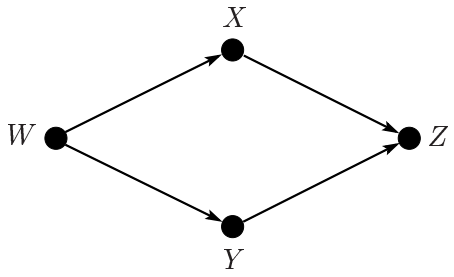
If X_S exhausts $\mathcal{X}_S$, i.e. all outcomes $x_S \in \mathcal{X}_S$ have a nonvanishing probability $p(x_S)$, then implication (7) *becomes an equivalence.*

Proof: Follows directly from the 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 the 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 the 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 following joint distribution:

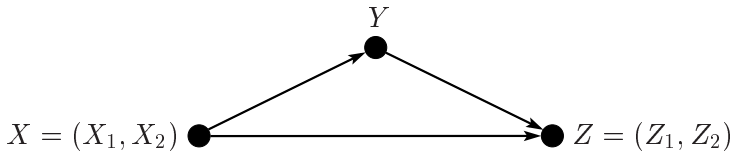
$$p(w,x,y,z) \;=\; \frac{1}{2}\,\delta_w(x)\,\delta_w(y)\,\delta_{\mathrm{XOR}(x,y)}(z)$$

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
$I_p(X:Y) = 1$	$I_p(X \to Y) = 0$
$I_p(X:Y \,\|\, W) = 0$	$I_p(X \to Y \,\|\, \widehat{W}) = 0$
$I_p(W:Z \,\|\, Y) = 0$	$I_p(W \to Z \,\|\, \widehat{Y}) = 1$

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 the 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 the nodes X and Y. For the node Z we consider two cases:
Case 1: Z is assumed to copy the state fom X, and we have the joint distribution

$$p(w,x,y,z) \;=\; \frac{1}{2}\,\delta_w(x)\,\delta_w(y)\,\delta_x(z) \tag{8}$$

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 machanism 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

$$p(w,x,y,z) \;=\; \frac{1}{2}\,\delta_w(x)\,\delta_w(y) \cdot \begin{cases} \delta_x(z), & \text{if } x = y \\ \frac{1}{2}\;, & \text{if } x \neq y \end{cases} \tag{9}$$

which coincides with the joint distribution (8) of the first case. But here, Y determines to some extent whether X can control the outcome of Z. More precisely, one has

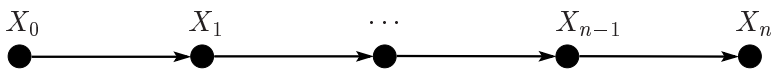
$$I_p(X \to Z\,|\,\widehat{Y}) \;=\; \frac{3}{4}\log\frac{4}{3} \;\approx\; 0.31$$

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

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



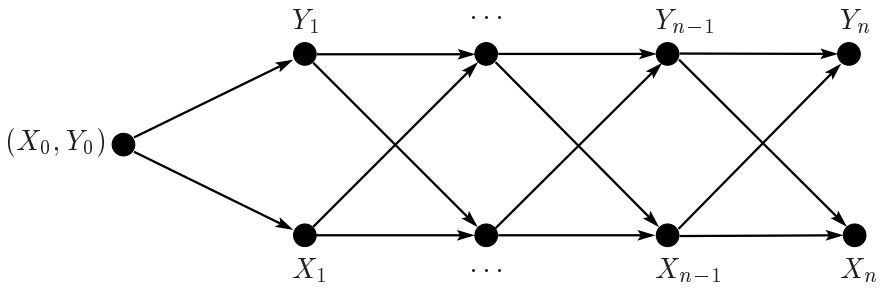
There is a field of research (Shaw, 1981, 1984; Matsumoto and Tsuda, 1988), which is not restricted to this simple setting, but also deals with more general dynamical systems, that aims at relating the qualitative characteristics of a given dynamics to its information flow in time. Hereby, information flow is usually quantified by the mutual information between a time interval $[i,j] = \{i, i+1, \ldots, j\}$ of the past and a time interval $[k,l] = \{k, k+1, \ldots, l\}$ of the future. Applied to our simple example, this would correspond to the mutual information

$$I_p(X_{[i,j]} : X_{[k,l]}), \qquad 1 \leq i \leq j < k \leq l \leq n \tag{10}$$

Within the context of the present paper, it is natural to ask whether our definition of information flow is consistent with the definition (10). Indeed, a small calculation proves

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

This consistency breaks down if one wants to quantify information flows among the elements of a composite dynamical system. To make clear in what sense this is meant we consider two processes X and Y as shown in the following figure:



The processes are assumed to be generated by an initial distribution p_0 and kernels p_X and p_Y as follows

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

Schreiber (Schreiber, 2000) 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 $1 \leq k < n$, it is defined as the conditional mutual information $I_p(Y_{[1,k]} : X_{k+1} \,|\, X_{[1,k]})$. The following simple but instructive example proves that the tranfer entropy does not necessarily coincide with the information flow $I_p(Y_{[1,k]} \to X_{k+1} \,|\, \widehat{X}_{[1,k]})$:

Example 6 (**Information Exchange**)**.** We consider two observationally equivalent cases:
Case 1: Assume that both nodes have states 0 and 1, and assume that at each time step k they just copy the state of the other node. If we start with a configuration (x, y) according to the distribution $\frac{1}{2}\left(\delta_{(0,1)} + \delta_{(1,0)}\right)$, we would observe a sequence $\cdots \to (0,1) \to (1,0) \to (0,1) \to \ldots$. The transfer entropy vanishes in this case for all times k. This contradicts the intuition that by copying the other node's state, clearly there is a flow of information. In consistence with this intuition, our measure of information flow has the maximal value of one bit in this case.
Case 2: Consider now the case that X_{k+1} is the inversion of X_k for all k (i.e. 0 becomes 1, and 1 becomes 0) and, likewise, Y_{k+1} is the inversion of Y_k. In particular, there is no interaction between X and Y after their initial generation. This is observationally equivalent to the first case and thus the transfer entropy remains 0. However, its interventional dynamics is different, and the information flow $I_p(Y_{[1,k]} \to X_{k+1} \,|\, \widehat{X}_{[1,k]})$ becomes 0 in this case. Thus information flow is able to distinguish the case of information being actively exchanged between the chains X and Y and the case where there is no such exchange.

In Examples 3 and 4 we imposed nodes lying between the "sender" and the "receiver" node. The examples show that imposing such nodes can both reduce (Example 4) or increase (Example 3) an information flow. The reduction of the flow by imposing intermediate nodes naturally fits intuition. However, the increase of the flow by imposing a node is a typical example of how the rules governing information flow differ from naive material flow. The fact that information flow can both increase or decrease by imposing nodes is closely related to the fact that *synergy* $I_p(X_A : X_B \,|\, X_S) - I_p(X_A : X_B)$ or triple mutual information quantities can be both positive and negative (Schneidman et al., 2003; Adami, 1998; Bell, 2003).

7. Application Scenarios

In Section 1, we have briefly mentioned some useful applications for the concept of information flow. The usefulness of the concept extends beyond that. We believe that the above measure of

the causal flow of information allows one to qunatify a number of phenomena. Here we wish to give a glimpse into possible perspectives for its future use.

Physics: Via the unambiguous causal interpretation of the information flow it is possible to enhance the identification of causal relations and mechanisms in general physical systems by a measure of their impact. This provides a new tool for quantitative studies of dynamical and complex systems. It would be interesting to pursue in how far above concept of information flow could be applied to the computational mechanics / causal states framework (Crutchfield and Young, 1989; Shalizi and Crutchfield, 2002).

Synchronization: Synchronization is a phenomenon of great interest in the context of self-organization (Strogatz, 2004). The information flow formalism can help elicit which information flows between the different components of a system are involved to create the effects of global synchronization.

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 (Sato and Ay, 2006; Sato et al., 2005). Here, one often encounters dynamics moving towards cooperative or antagonistic player behaviour. Using information flow would allow one to attribute how much a given player is "responsible" for the emergence of a particular cooperative or antagonistic outcome.

Models for the Perception-Action Loop: In Section 1 some work using information flow-type quantities has been briefly mentioned. Information-theoretic principles, long believed to be relevant for the understanding of biological information processing (Barlow, 1959; Atick, 1992) now begin to receive renewed attention (Linsker, 1988; Baddeley et al., 2000). Related to that, Bayesian and prediction-based concepts of the self-organization of the perception-action loop prove themselves increasingly successful (Körding and Wolpert, 2004; Der et al., 1999; Porr et al., 2006). The family of information flow methods thus promises to provide a calculus by which principles guiding biological (and artificial) perception-action loops can be identified and formulated (Klyubin et al., 2004).

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

8. Conclusions

The present 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, while, on the other hand, realizing 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 requires us to take into account the causal nature of the systems under study. For this, we used the interventional formalism from (Pearl, 2000) which provided an appropriate framework for the causal mechanisms in the given system. The classical mutual information can be introduced by quantifying the deviation of two random variables from stochastic independence. Analogously, we introduced information flow as the deviation of two random variables from causal independence by appropriately adapting the quantities involved in establishing probabilistic independence.

In a number of examples we have shown that our measure for information flow is indeed different from other 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).

Together with information flow, we have developed an appropriate modification of well-established formalisms to fit the framework of causal Bayesian networks. Thus, we have shown how the notion of information flow comes together with a broad and robust set of conceptual tools.

The concept of causality and information flow shows nicely how the possibility for intervention (or "experiment") modifies our understanding about 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 seems more intuitive than the observational concept of *d*-separation; this is consistent with Pearl's philosophy insofar as causal knowledge seems to be less brittle than observational (probabilistic) knowledge (Pearl, 2000).

New notions are typically introduced as generalizations or adaptations of existing and established concepts, driven by theoretical considerations. However, one of the strongest justifications for introducing a new notion is the practical need for a notion with suitable properties. This exactly was the case for information flow. If well constructed, such a notion can not just help covering the cases that motivated its introduction, but also open up pathways towards 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 to achieve 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 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

Adami, C. (1998). *Introduction to Artificial Life*. Springer.

Atick, J. J. (1992). Could Information Theory Provide an Ecological Theory of Sensory Processing. *Network: Computation in Neural Systems*, 3(2):213–251.

Ay, N. and Krakauer, D. C. (2006). Information geometric theories for robust biological networks. *J. Theor. Biology*. In Press.

Baddeley, R., Hancock, P., and Földiák, P., editors (2000). *Information Theory and the Brain*. Cambridge University Press.

Barlow, H. B. (1959). Possible Principles Underlying the Transformations of Sensory Messages. In Rosenblith, W. A., editor, *Sensory Communication: Contributions to the Symposium on Principles of Sensory Communication*, pages 217–234. The M.I.T. Press.

Bell, A. J. (2003). The co-information lattice. In Amari, S., Cichocki, A., Makino, S., and Murata, N., editors, *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA 2003*.

Crutchfield, J. P. and Young, K. (1989). Inferring statistical complexity. *Phys. Rev. Lett.*, 63:105–108.

Der, R., Steinmetz, U., and Pasemann, F. (1999). Homeokinesis – A new principle to back up evolution with learning. In Mohammadian, M., editor, *Computational Intelligence for Modelling, Control, and Automation*, volume 55 of *Concurrent Systems Engineering Series*, pages 43–47. IOS Press.

Klyubin, A., Polani, D., and Nehaniv, C. (2006). In preparation.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2004). Organization of the Information Flow in the Perception-Action Loop of Evolved Agents. In *Proceedings of 2004 NASA/DoD Conference on Evolvable Hardware*, pages 177–180. IEEE Computer Society.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005). Empowerment: A Universal Agent-Centric Measure of Control. In *Proc. IEEE Congress on Evolutionary Computation, 2-5 September 2005, Edinburgh, Scotland (CEC 2005)*, pages 128–135. IEEE.

Körding, K. P. and Wolpert, D. M. (2004). Bayesian Integration in Sensorimotor Learning. *Nature*, 427:244–247.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.

Lauritzen, S. L. (2005). *Graphical Models for Causal Inference*. Royal Economics Society Summer School, Oxford. Lecture Notes.

Linsker, R. (1988). Self-Organization in a Perceptual Network. *Computer*, 21(3):105–117.

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

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

Porr, B., Egertony, A., and Wörrgötter, F. (2006). Towards Closed Loop Information: Predictive Information. *Constructivist Foundations*, 1(2).

Sato, Y., Akiyama, E., and Crutchfield, J. P. (2005). Stability and Diversity in Collective Adaptation. *Physica D*, 210:21–57.

Sato, Y. and Ay, N. (2006). Adaptive dynamics of interacting Markovian processes. in preparation.

Schneidman, E., Bialek, W., and Berry II, M. J. (2003). Synergy, Redundancy, and Independence in Population Codes. *The Journal of Neuroscience*, 23(37):11539–11553.

Schreiber, T. (2000). Measuring Information Transfer. *Phys. Rev. Lett.*, 85:461–464.

Shalizi, C. R. and Crutchfield, J. P. (2002). Information Bottlenecks, Causal States, and Statistical Relevance Bases: How to Represent Relevant Information in Memoryless Transduction. *Advances in Complex Systems*, 5(1):91–95.

Shaw, R. (1981). Strange attractors, Chaotic behavior and information flow. *Zeitschrift für Naturforschung*, 36:80.

Shaw, R. (1984). *The dripping faucet as a model chaotic system*. Aerial Press, Santa Cruz, CA.

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

Wennekers, T. and Ay, N. (2005). Finite State Automata Resulting From Temporal Information Maximization. *Neural Computation*, 17(10):2258–2290.

Werner, A., Hötzl, H., Käss, W., and Maloszewski, P. (1997). Interpretations of Tracer Experiments in the Danube-Aach-System, Western Swabian Alb, Germany, with analytical models. In Günay and Johnson, editors, *Karst Waters & Environmental Impacts*, pages 153–160, Rotterdam. Balkema.

Wheeler, J. A. (1990). Information, Physics, Quantum: The Search for Links. In Zurek, W. H., editor, *Complexity, Entropy and the Physics of Information*, Santa Fe Studies in the Sciences of Complexity, pages 3–28, Reading, Mass. Addison-Wesley.

E-mail address: nay@mis.mpg.de

E-mail address: d.polani@herts.ac.uk