The Annals of Statistics
2013, Vol. 41, No. 5, 2324–2358
DOI: 10.1214/13-AOS1145

QUANTIFYING CAUSAL INFLUENCES

BY DOMINIK JANZING[*], DAVID BALDUZZI[*,†], MORITZ GROSSE-WENTRUP[*] AND BERNHARD SCHÖLKOPF[*]

Max Planck Institute for Intelligent Systems[*] *and ETH Zürich*[†]

Many methods for causal inference generate directed acyclic graphs (DAGs) that formalize causal relations between n variables. Given the joint distribution on all these variables, the DAG contains all information about how intervening on one variable changes the distribution of the other $n-1$ variables. However, *quantifying* the causal influence of one variable on another one remains a nontrivial question.

Here we propose a set of natural, intuitive postulates that a measure of causal strength should satisfy. We then introduce a communication scenario, where edges in a DAG play the role of channels that can be locally corrupted by interventions. Causal strength is then the relative entropy distance between the old and the new distribution.

Many other measures of causal strength have been proposed, including average causal effect, transfer entropy, directed information, and information flow. We explain how they fail to satisfy the postulates on simple DAGs of ≤ 3 nodes. Finally, we investigate the behavior of our measure on time-series, supporting our claims with experiments on simulated data.

1. Introduction. Inferring causal relations is among the most important scientific goals since causality, as opposed to mere statistical dependencies, provides the basis for reasonable human decisions. During the past decade, it has become popular to phrase causal relations in directed acyclic graphs (DAGs) [14] with random variables (formalizing statistical quantities after repeated observations) as nodes and causal influences as arrows.

We briefly explain this formal setting. Here and throughout the paper, we assume causal sufficiency, that is, there are no hidden variables that influence more than one of the n observed variables. Let G be a causal DAG with nodes $X_1, \ldots, X_n$ where $X_i \to X_j$ means that X_i influences X_j "directly" in the sense that intervening on X_i changes the distribution of X_j even if all other variables are held constant (also by interventions). To simplify notation, we will mostly assume the X_j to be discrete. $P(x_1, \ldots, x_n)$ denotes the probability mass function of the joint distribution $P(X_1, \ldots, X_n)$. According to the Causal Markov Condition [14, 21], which we take for granted in this paper, every node X_j is conditionally inde-

Received August 2012; revised June 2013.
MSC2010 subject classifications. 62-09, 62M10.
Key words and phrases. Causality, Bayesian networks, information flow, transfer entropy.

pendent of its nondescendants, given its parents with respect to the causal DAG G. If PA_j denotes the set of parent variables of X_j (i.e., its direct causes) in G, the joint probability thus factorizes [10] into

$$P(x_1, \dots, x_n) = \prod_{j=1}^{n} P(x_j | pa_j), \tag{1}$$

where pa_j denotes the values of PA_j. By slightly abusing the notion of conditional probabilities, we assume that $P(X_j|pa_j)$ is also defined for those pa_j with $P(pa_j) = 0$. In other words, we know how the causal mechanisms act on potential combinations of values of the parents that never occur. Note that this assumption has implications because such causal conditionals cannot be learned from observational data even if the causal DAG is known.

Given this formalism, why define causal strength? After all, the DAG together with the causal conditionals contain the complete causal information: one can easily compute how the joint distribution changes when an external intervention sets some of the variables to specific values [14]. However, describing causal relations in nature with a DAG always requires first deciding how detailed the description should be. Depending on the desired precision, one may want to account for some weak causal links or not. Thus, an objective measure distinguishing weak arrows from strong ones is required.

1.1. *Related work.* We discuss some definitions of causal strength that are either known or just come up as straightforward ideas.

Average causal effect: Following [14], $P(Y|do(X = x))$ denotes the distribution of Y when X is set to the value x [it will be introduced more formally in equation (6)]. Note that it only coincides with the usual conditional distribution $P(Y|x)$ if the statistical dependence between X and Y is due to a direct influence of X on Y, with no confounding common cause. If all X_i are binary variables, causal strength can then be quantified by the Average Causal Effect [7, 14]

$$\mathrm{ACE}(X_i \to X_j) := P\big(X_j = 1|do(X_i = 1)\big) - P\big(X_j = 1|do(X_i = 0)\big).$$

If a real-valued variable X_j is affected by a binary variable X_i, one considers the shift of the mean of X_j that is caused by switching X_i from 0 to 1. Formally, one considers the difference [13]

$$\mathbb{E}\big(X_j|do(X_i = 1)\big) - \mathbb{E}\big(X_j|do(X_i = 0)\big).$$

This measure only accounts for the linear aspect of an interaction since it does not reflect whether X_i changes higher order moments of the distribution of X_j.

Analysis of Variance (*ANOVA*): Let X_i be caused by $X_1, \dots, X_{i-1}$. The variance of X_i can formally be split into the average of the variances of X_i, given X_k with $k \leq i - 1$, and the variance of the expectations of X_i, given X_k:

$$\mathrm{Var}(X_i) = \mathbb{E}\big(\mathrm{Var}(X_i|X_k)\big) + \mathrm{Var}\big(\mathbb{E}(X_i|X_k)\big). \tag{2}$$

In the common scenario of drug testing experiments, for instance, the first term in equation (2) is given by the variability of X_i within a group of equal treatments (i.e., fixed x_k), while the second one describes how much the means of X_i vary between different treatments. It is tempting to say that the latter describes the part of the total variation of X_i that is *caused by* the variation of X_k, but this is conceptually wrong for nonlinear influences and if there are statistical dependencies between X_k and the other parents of X_i [11, 13].

For linear structural equations,

$$X_i = \sum_{j<i} \alpha_{ij} X_j + E_i \qquad \text{with } E_j \text{ being jointly independent,}$$

and additionally assuming X_k to be independent of the other parents of X_i, the second term is given by $\mathrm{Var}(\alpha_{ik} X_k)$, which indeed describes the amount by which the variance of X_i decreases when X_k is set to a fixed value by intervention. In this sense,

$$r_{ik} := \frac{\mathrm{Var}(\alpha_{ik} X_k)}{\mathrm{Var}(X_i)} \tag{3}$$

is indeed the fraction of the variance of X_i that is *caused by* X_k. By rescaling all X_j such that $\mathrm{Var}(X_j) = 1$, we have $r_{ik} = \alpha_{ik}^2$. Then, the square of the structure coefficients itself can be seen as a simple measure for causal strength.

(*Conditional*) *Mutual information*: The information of X on Y or vice versa is given by [5]

$$I(X;Y) := \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}.$$

The information of X on Y or vice versa if Z is given is defined by [5]

$$I(X;Y|Z) := \sum_{x,y,z} P(x,y,z) \log \frac{P(x,y|z)}{P(x|z)P(y|z)}. \tag{4}$$

There are situations where these expressions (with Z describing some background condition) can indeed be interpreted as measuring the strength of the arrow $X \to Y$. An essential part of this paper describes the conditions where this makes sense and how to replace the expressions with other information-theoretic ones when it does not.

Granger causality/Transfer entropy/Directed information: Quantifying causal influence between time series [e.g. between $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$] is special because one is interested in quantifying the effect of all (X_t) on all (Y_{t+s}). If we represent the causal relations by a DAG where every time instant defines a separate pair of variables, then we ask for the strength of a *set of arrows*. If X_t and Y_t are considered as instances of the variables X, Y, we leave the regime of i.i.d. sampling.

Measuring the reduction of uncertainty in one variable after knowing another is also a key idea in several related methods for quantifying causal strength in time series. Granger causality in its original formulation uses reduction of variance [6]. Nonlinear information-theoretic extensions in the same spirit are transfer entropy [20] and directed information [12]. Both are essentially based on conditional mutual information, where each variable X, Y, Z in (4) is replaced with an appropriate set of variables.

Information flow: Since the above measures quantify dependencies rather than causality, several authors have defined causal strength by replacing the observed probability distribution with distributions that arise after interventions (computed via the causal DAG). [3] defined Information Flow via an operation, "source exclusion", which removes the influence of a variable in a network. [4] defined a different notion of Information Flow explicitly via Pearl's *do*-calculus. Both measures are close to ours in spirit and in fact the version in [3] coincides with ours when quantifying the strength of a single arrow. However, both do not satisfy our postulates.

Mediation analysis: [2, 15, 19] explore how to separate the influence of X on Y into parts that can be attributed to specific paths by "blocking" other paths. Consider, for instance, the case where X influences Y directly and indirectly via $X \to Z \to Y$. To test its direct influence, one changes X from some "reference" value x' to an "active" value x while keeping the distribution of Z that either corresponds to the reference value x' or to the natural distribution $P(X)$. A natural distinction between a reference state and an active state occurs, for instance, in drug testing scenario where taking the drug means switching from reference to active. In contrast, our goal is not to study the impact of one specific switching from x' to x. Instead, we want to construct a measure that quantifies the direct effect of the variable X on Y, while treating all possible values of X in the same way. Nevertheless, there are interesting relation between these approaches and ours that we briefly discuss at the end of Section 4.2.

2. Postulates for causal strength.

Let us first discuss the properties we expect a measure of causal strength to have. The key idea is that causal strength is supposed to measure the impact of an intervention that removes the respective arrows. We present five properties that we consider reasonable. Let $\mathfrak{C}_S$ denote the strength of the arrows in set S. By slightly overloading notation, we write $\mathfrak{C}_{X\to Y}$ instead of $\mathfrak{C}_{\{X\to Y\}}$.

P0. *Causal Markov condition*: If $\mathfrak{C}_S = 0$, then the joint distribution satisfies the Markov condition with respect to the DAG G_S obtained by removing the arrows in S.

P1. *Mutual information*: If the true causal DAG reads $X \to Y$, then

$$\mathfrak{C}_{X\to Y} = I(X; Y).$$

P2. *Locality*: The strength of $X \to Y$ only depends on (1) how Y depends on X and its other parents, and (2) the joint distribution of all parents of Y. Formally, knowing $P(Y|PA_Y)$ and $P(PA_Y)$ is sufficient to compute $\mathfrak{C}_{X\to Y}$. For strictly positive densities, this is equivalent to knowing $P(Y, PA_Y)$.

P3. *Quantitative causal Markov condition*: If there is an arrow from X to Y, then the causal influence of X on Y is greater than or equal to the conditional mutual information between Y and X given all the other parents of Y. Formally

$$\mathfrak{C}_{X\to Y} \geq I\big(X; Y|PA_Y^X\big).$$

P4. *Heredity*: If the causal influence of a set of arrows is zero, then the causal influence of all its subsets (in particular, individual arrows) is also zero.

$$\text{If } S \subset T, \text{ then } \mathfrak{C}_T = 0 \Longrightarrow \mathfrak{C}_S = 0.$$

Note that we do not claim that *every* reasonable measure of causal strength should satisfy these postulates, but we now explain why we consider them natural and show that the postulates make sense for simple DAGs.

P0: If the purpose of our measure of causal strength is to quantify relevance of arrows, then removing a set of arrows with zero strength must make no difference. If, for instance, $\mathfrak{C}_{X\to Y} = 0$, removing $X \to Y$ should not yield a DAG that is ruled out by the causal Markov condition.

We should emphasize that $\mathfrak{C}_S$ can be nonzero even if S consists of arrows each individually having zero strength.

P1: The mutual information actually measures the strength of statistical dependencies. Since all these dependencies are generated by the influence of X on Y (and not by a common cause or Y influencing X), it makes sense to measure causal strength by strength of dependencies. Note that mutual information $I(X;Y) = H(Y) - H(Y|X)$ also quantifies the variability in Y that is due to the variability in X, see also Section A.4.

Mutual information versus channel capacity. Given the premise that causal strength should be an information-like quantity, a natural alternative to mutual information is the capacity of the information channel $x \mapsto P(Y|x)$, that is, the maximum over all values of mutual information $I_{Q(X)}(X;Y)$ for all input distributions $Q(X)$ of X when keeping the conditional $P(Y|X)$.

While mutual information $I(X;Y)$ quantifies the observable dependencies, channel capacity quantifies the strength of the strongest dependencies that can be generated using the information channel $P(Y|X)$. In this sense, $I(X;Y)$ quantifies the *factual* causal influence, while channel capacity measures the *potential* influence. Channel capacity also accounts for the impact of setting x to values that rarely or never occur in the observations. However, this sensitivity regarding effects of rare inputs can certainly be a problem for estimating the effect from sparse data. We therefore prefer mutual information $I(X;Y)$ as it better assesses the extent to which *frequently observed changes* in X influence Y.

P2: Locality implies that we can ignore causes of X when computing $\mathfrak{C}_{X\to Y}$, unless they are at the same time direct causes of Y. Likewise, other effects of Y are irrelevant. Moreover, it does not matter *how* the dependencies between the parents are generated (which parent influences which one or whether they are effects of a common cause), we only need to know their joint distribution with X.

Violations of locality have paradoxical implications. Assume, for example, variable Z would be relevant in DAG 1(a). Then, $\mathfrak{C}_{X\to Y}$ would depend on the mechanism that generates the distribution of X, while we are actually concerned with the information flowing from X to Y instead of that flowing *to* X from other nodes. Likewise, [see DAGs 1(b) and 1(c)] it is irrelevant whether X and Y have further effects.

P3: To justify the name of this postulate, observe that the restriction of P0 to the single arrow case $S = \{X \to Y\}$ is equivalent to

$$\mathfrak{C}_{X\to Y} = 0 \quad \Longrightarrow \quad I\big(Y; X | PA_Y^X\big) = 0.$$

To see this, we use the ordered Markov condition [14], Theorem 1.2.6, which is known to be equivalent to the Markov condition mentioned in the Introduction. It states that every node is conditionally independent of its predecessors (according to some ordering consistent with the DAG), given its parents. If PR_Y denotes the predecessors of Y for some ordering that is consistent with G and G_S, the ordered Markov condition for G_S holds iff

$$Y \perp\!\!\!\perp PR_Y | PA_Y^X, \tag{5}$$

since the conditions for all other nodes remain the same as in G. Due to the semi-graphoid axioms (weak union and contraction rule [14]), (5) is equivalent to

$$Y \perp\!\!\!\perp PR_Y \setminus \{X\} | PA_Y \wedge Y \perp\!\!\!\perp X | PA_Y^X.$$

Since the condition on the left is guaranteed by the Markov condition on G, the Markov condition on G_S is equivalent to $I(Y; X|PA_Y^X) = 0$.

In words, the arrow $X \to Y$ is the only reason for the conditional dependence $I(Y; X|PA_Y^X)$ to be nonzero, hence it is natural to postulate that its strength cannot be smaller than the dependence that it generates. Section 4.3 explains why we should not postulate equality.

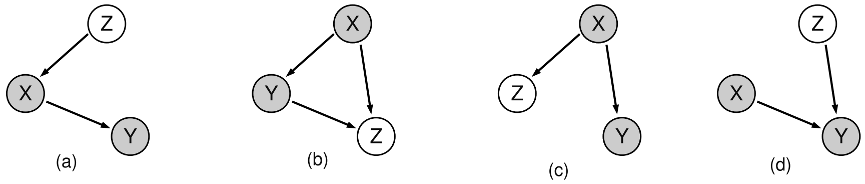

FIG. 1. *DAGs for which the* (*conditional*) *mutual information is a reasonable measure of causal strength*: *For* (a) *to* (c), *our postulates imply* $\mathfrak{C}_{X\to Y} = I(X;Y)$. *For* (d) *we will obtain* $\mathfrak{C}_{X\to Y} = I(X;Y|Z)$. *The nodes* X *and* Y *are shaded because they are source and target of the arrow* $X \to Y$, *respectively.*

P4: The postulate provides a compatibility condition: if a set of arrows has zero causal influence, and so can be eliminated without affecting the causal DAG, then the same should hold for all subsets of that set. We refer to this as the heredity property by analogy with matroid theory, where heredity implies that every subset of an independent set is independent.

3. Problems of known definitions.

Our definition of causal strength is presented in Section 4. This section discusses problems with alternate measures of causal strength.

3.1. *ACE and ANOVA*.

The first two measures are ruled out by P0. Consider a relation between three binary variables X, Y, Z, where $Y = X \oplus Z$ with X and Z being unbiased and independent. Then changing X has no influence on the statistics of Y. Likewise, knowing X does not reduce the variance of Y. To satisfy P0, we need modifications that account for the fact that we do observe an influence of X on Y for each fixed value z although this influence becomes invisible after marginalizing over Z.

3.2. *Mutual information and conditional mutual information*.

It suffices to consider a few simple DAGs to illustrate why mutual information and conditional mutual information are *not* suitable measures of causal strength in general.

Mutual information is not suitable in Figure 2(a). It is clear that $I(X;Y)$ is inappropriate because we can obtain $I(X;Y) \neq 0$ even when the arrow $X \to Y$ is missing, due to the common cause Z.

Conditional mutual information is not suitable for Figure 2(a). Consider the limiting case where the direct influence $Z \to Y$ gets weaker until it almost disappears ($P(y|x,z) \approx P(y|x)$). Then the behavior of the system (observationally and interventionally) is approximately described by the DAG 1(a). Using $I(X;Y|Z)$ makes no sense in this scenario since, for example, X may be obtained from Z by a simple copy operation, in which case $I(X;Y|Z) = 0$ necessarily, even when X influences Y strongly.

3.3. *Transfer entropy*.

Transfer entropy [20] is intended to measure the influence of one time-series on another one. Let $(X_t, Y_t)_{t \in Z}$ be a bivariate stochastic

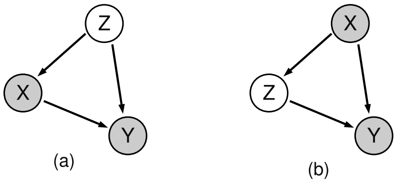

FIG. 2. *DAGs for which finding a proper definition of $\mathfrak{C}_{X \to Y}$ is challenging.*

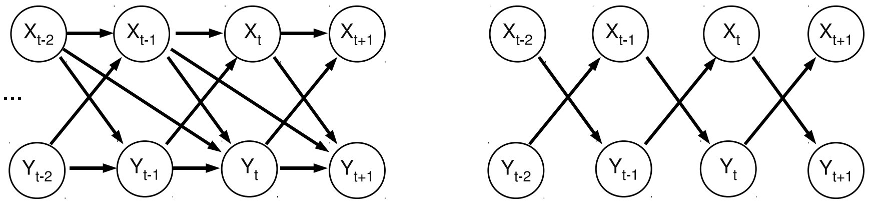

FIG. 3. *Left*: *Typical causal DAG for two time series with mutual causal influence. The structure is acyclic because instantaneous influences are excluded. Right*: *counter example in* [4]. *Transfer entropy vanishes if all arrows are copy operations although the time series strongly influence each other.*

process where X_t influence some Y_s with $s > t$, see Figure 3, left. Then transfer entropy is defined as the following conditional mutual information:

$$I(X_{(-\infty,t-1]} \rightarrow Y_t | Y_{(-\infty,t-1]}) := I(X_{(-\infty,t-1]}; Y_t | Y_{(-\infty,t-1]}).$$

It measures the amount of information the past of X provides about the present of Y given the past of Y. To quantify causal influence by conditional information relevance is also in the spirit of Granger causality, where information is usually understood in the sense of the amount of reduction of the linear prediction error.

Transfer entropy is an unsatisfactory measure of causal strength. [4] pointed out that transfer entropy fails to quantify causal influence for the following toy model: Assume the information from X_t is perfectly copied to Y_{t+1} and the information from Y_t to X_{t+1} (see Figure 3, right). Then the past of Y is already sufficient to perfectly predict the present value of Y and the past of X does not provide any further information. Therefore, transfer entropy vanishes although both variables heavily influence one another. If the copy operation is noisy, transfer entropy is nonzero and thus seems more reasonable, but the quantitative behavior is still wrong (as we will argue in Example 7).

Transfer entropy violates the postulates. Transfer entropy yields 0 bits of causal influence in a situation where common sense and P1 together with P2 require that causal strength is 1 bit (P2 reduces the DAG to one in which P1 applies). Since our postulates refer to the strength of a *single* arrow while transfer entropy is supposed to measure the strength of all arrows from X to Y, we reduce the DAG such that there is only one arrow from X to Y; see Figure 4. Then,

$$\begin{aligned} I(X_{(-\infty,t-1]} \rightarrow Y_t | Y_{(-\infty,t-1]}) &= I(X_{(-\infty,t-1]}; Y_t | Y_{(-\infty,t-1]}) \\ &= I(X_{t-1}; Y_t | Y_{t-2}). \end{aligned}$$

The causal structure coincides with DAG 1(a) by setting $Y_{t-2} \equiv Z$, $X_{t-1} \equiv X$, and $Y_t \equiv Y$. With these replacements, transfer entropy yields $I(X; Y|Z) = 0$ bits instead of $I(X; Y) = 1$ bit, as required by P1 and P2.

Note that the same problem occurs if causal strength between time series is quantified by directed information [12] because this measure also conditions on the entire past of Y.

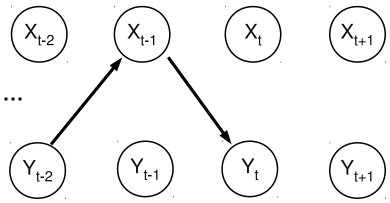

FIG. 4. *Time series with only two causal arrows, where transfer entropy fails satisfying our postulates.*

3.4. *Information flow.* Note that [4] and [3] introduce two different quantities, both called "information flow." We consider them in turn.

After arguing that transfer entropy does not properly capture the strength of the impact of interventions, [4] proposes to define causal strength using Pearl's *do* calculus [14]. Given a causal directed acyclic graph G, Pearl computes the joint distribution obtained if variable X_j is forcibly set to the value x_j as

$$P(x_1, \dots, x_n | do(x'_j)) := \prod_{i \neq j} P(x_i | pa_i) \cdot \delta_{x_j, x'_j}. \tag{6}$$

Intuitively, the intervention on X_j removes the dependence of X_j on its parents and therefore replaces $P(x_j | pa_j)$ with the kronecker symbol. Likewise, one can define interventions on several nodes by replacing all conditionals with kronecker symbols.

Given three sets of nodes X_A, X_B and X_C in a directed acyclic graph G, information flow is defined by

$$\begin{aligned} I(X_A \to X_B | do(X_C)) \\ := \sum_{x_C, x_A, x_B} P(x_C) P(x_A | do(x_C)) P(x_B | do(x_A, x_C)) \\ \times \log \frac{P(x_B | do(x_A, x_C))}{\sum_{x'_A} P(x'_A | do(x_C)) P(x_B | do(x'_A, x_C))}. \end{aligned}$$

To better understand this expression, we first consider the case where the set X_C is empty. Then we obtain

$$I(X_A \to X_B) := \sum_{x_A, x_B} P(x_A) P(x_B | do(x_A)) \log \frac{P(x_B | do(x_A))}{\sum_{x'_A} P(x'_A) P(x_B | do(x'_A))},$$

which measures the mutual information between X_A and X_B obtained when the information channel $x_A \mapsto P(X_B | do(x_A))$ is used with the input distribution $P(X_A)$.

Information flow, as defined in [4], *is an unsatisfactory measure of causal strength.* To quantify $X \to Y$ in DAGs 2(a) and 2(b) using information flow, we may either choose $I(X \to Y)$ or $I(X \to Y | do(Z))$. Both choices are inconsistent with our postulates and intuitive expectations.

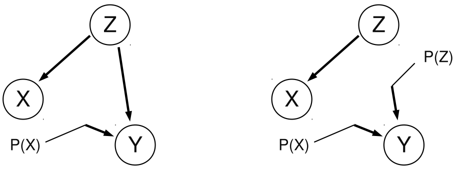

FIG. 5. *Left*: *deletion of the arrow* $X \to Y$. *The conditional* $P(Y|X,Z)$ *is fed with an independent copy of* X, *distributed with* $P(X)$. *The resulting distribution reads* $P_{X\to Y}(x,y,z) = P(x,z)\sum_{x'} P(y|z,x')P(x')$. *Right*: *deletion of both incoming arrows. The conditional* $P(Y|X,Z)$ *is then fed with the product distribution* $P(X)P(Z)$ *instead of the joint* $P(X,Z)$ *as in* [3], *since the latter would require communication between the open ends. We obtain* $P_{X\to Y,Z\to Y}(x,y,z) = \sum_{x',z'} P(x,z)P(y|x',z')P(x')P(z')$. *Feeding with independent inputs is particularly relevant for the following example*: *let X and Z be binary with* $X = Z$ *and* $Y = X \oplus Z$. *Then*, *the cutting had no impact if we would keep the dependences*.

Start with $I(X \to Y)$ and DAG 2(a). Let X, Y, Z be binary with $Y := X \oplus Z$ an XOR. Let Z be an unbiased coin toss and X obtained from Z by a faulty copy operation with two-sided symmetric error. One easily checks that $I(X \to Y)$ is zero in the limit of error probability $1/2$ (making X and Y independent). Nevertheless, dropping the arrow $X \to Y$ violates the Markov condition, contradicting P0. For error rate close to $1/2$, we still violate P3 because $I(Y; X|Z)$ is close to 1, while $I(X \to Y)$ is close to zero. A similar argument applies to DAG 2(b).

Now consider $I(X \to Y|do(Z))$. Note that it yields different results for DAGs 2(a) and 2(b) when the joint distribution is the same, contradicting P2. This is because $P(x|do(z)) = P(x|z)$ for 2(a), while $P(x|do(z)) = P(x)$ for 2(b). In other words, $I(X \to Y|do(Z))$ depends on the causal relation *between* the two causes X and Z, rather than only on the relation between causes and effects.

Apart from being inconsistent with our postulate, it is unsatisfactory that $I(X \to Y|do(Z))$ tends to zero for the example above if the error rate of copying X from Z in DAG 2(a) tends to zero (conditioned on setting Z to some value, the information passed from X to Y is zero because X attains a fixed value, too). In this limit, Y is always zero. Clearly, however, link $X \to Y$ is important for explaining the behavior of the XOR: without the link, the gate would not output "zero" for both $Z = 0$ and $Z = 1$.

Information flow, as defined in [3], *is unsatisfactory as a measure of causal strength for sets of edges.* Since this measure is close to ours, we will explain (see caption of Figure 5) the difference when introducing ours and show that P4 fails without our modification.

4. Defining the strength of causal arrows.

4.1. *Definition in terms of conditional probabilities.* This section proposes a way to quantify the causal influence of a set of arrows that yields satisfactory answers in all the cases discussed above. Our measure is motivated by a scenario where nodes represent different parties communicating with each other via channels. Hence, we think of arrows as physical channels that propagate information between distant points in space, for example, wires that connect electronic devices. Each such wire connects the output of a device with the input of another one. For the intuitive ideas below, it is also important that the wire connecting X_i and X_j physically contains full information about X_i [which may be more than the information that is required to explain the output behavior $P(X_j|PA_j)$]. We then think of the strength of arrow $X_i \to X_j$ as the impact of corrupting it, that is, the impact of cutting the wire. To get a well-defined "post-cutting" distribution we have to say what to do with the open end corresponding to X_j, because it needs to be fed with some input. It is natural to feed it probabilistically with inputs x_i according to $P(X_i)$ because this is the only distribution of X_i that is locally observable [feeding it with some conditional distribution $P(X_i|\cdot\cdot)$ assumes that the one cutting the edge has access to other nodes—and not only the physical state of the channel]. Note that this notion of cutting edges coincides with the "source exclusion" defined in [3] if only one edge is cut. However, we define the deletion of a *set* of arrows by feeding all open ends with the product of the corresponding marginal distributions, while [3] keeps the dependencies between the open ends and removes the dependencies between open ends and the other variables. Our post-cutting distribution can be thought of as arising from a scenario where each channel is cut by an independent attacker, who tries to blur the attack by feeding her open end with $P(X_i)$ (which is the only distribution she can see), while [3] requires communicating attackers who agree on feeding their open ends with the observed joint distribution.

Lemma 1 and Remark 1 below provide a more mathematical argument for the product distribution. Figure 5 visualizes the deletion of one edge (left) and two edges (right).

We now define the "post-cutting" distribution formally:

DEFINITION 1 (Removing causal arrows). Let G be a causal DAG and P be Markovian with respect to G. Let $S \subset G$ be a set of arrows. Set PA_j^S as the set of those parents X_i of X_j for which $(i, j) \in S$ and $PA_j^{\bar{S}}$ those for which $(i, j) \notin S$. Set

$$P_S(x_j|pa_j^{\bar{S}}) := \sum_{pa_j^S} P(x_j|pa_j^{\bar{S}}, pa_j^S) P_{\prod}(pa_j^S), \tag{7}$$

where $P_{\prod}(pa_j^S)$ denotes for a given j the product of marginal distributions of all variables in PA_j^S. Define a new joint distribution, the *interventional distribution*[1]

$$P_S(x_1, \dots, x_n) := \prod_j P_S(x_j|pa_j^{\bar{S}}). \tag{8}$$

[1] Note that this intervention differs from the kind of interventions considered by [14], where variables are set to specific values. Here we intervene on the arrows, the "information channels," and not on the nodes.

See Figure 5, left, for a simple example with cutting only one edge. Equation (8) formalizes the fact that each open end of the wires is independently fed with the corresponding marginal distribution, see also Figure 5, right. Information flow in the sense of [3] is obtained when the product distribution $P_\Pi(pa_j^{\bar{S}})$ in (7) is replaced with the joint distribution $P(pa_j^{\bar{S}})$.

The modified joint distribution P_S can be considered as generated by the reduced DAG:

LEMMA 1 (Markovian). *The interventional distribution P_S is Markovian with respect to the graph G_S obtained from G by removing the edges in S.*

PROOF. By construction, P_S factorizes according to G_S in the sense of (1). □

REMARK 1. Markovianity is violated if the dependencies between open ends are kept. Consider, for instance, the DAG $X \to Y \to Z$. Cutting both edges yields

$$P_S(x, y, z) = P(x) \sum_{x'} P(y|x') P(x') \sum_{y'} P(z|y') P(y') = P(x) P(y) P(z),$$

which is obviously Markovian with respect to the DAG without arrows. Feeding the "open ends" with $P(x', y')$ instead obtains

$$\tilde{P}_S(x, y, z) = P(x) \sum_{x'y'} P(y|x') P(z|y') P(x', y'),$$

which induces dependencies between Y and Z, although we have claimed to have removed all links between the three variables.

DEFINITION 2 (Causal influence of a set of arrows). The causal influence of the arrows in S is given by the Kullback–Leibler divergence

$$\mathfrak{C}_S(P) := D(P \| P_S). \tag{9}$$

If $S = \{X_k \to X_l\}$ is a single edge we write $\mathfrak{C}_{k\to l}$ instead of $\mathfrak{C}_{X_k \to X_l}$.

REMARK 2 (Observing versus intervening). Note that P_S could easily be confused with a different distribution obtained when the open ends are fed with conditional distributions rather than marginal distributions. As an illustrative example, consider DAG 2(a) and define $\tilde{P}_{X\to Y}(X, Y, Z)$ as

$$\tilde{P}_{X\to Y}(x, y, z) := P(x, z) P(y|z) = P(x, z) \sum_{x'} P(y|x') P(x'|z),$$

and recall that replacing $P(x'|z)$ with $P(x')$ in the right most expression yields $P_{X\to Y}$. We call $\tilde{P}_{X\to Y}$ the "partially observed distribution." It is the distribution obtained by ignoring the influence of X on Y: $\tilde{P}_{X\to Y}$ is computed according to (1), but uses a DAG where $X \to Y$ is missing. The difference between "ignoring" and "cutting" the edge is important for the following reason. By a known rephrasing of mutual information as relative entropy [5] we obtain

$$D(P\|\tilde{P}_{X\to Y}) = \sum_{x,y,z} P(x,y,z)\log\frac{P(y|z,x)}{P(y|z)} = I(X;Y|Z), \tag{10}$$

which, as we have already discussed, is *not* a satisfactory measure of causal strength. On the other hand, we have

$$\mathfrak{C}_{X\to Y} = D(P\|P_{X\to Y}) = D\big[P(Y|Z,X)\|P_{X\to Y}(Y|Z,X)\big] \tag{11}$$

$$\begin{aligned} &= D\big[P(Y|Z,X)\|P_{X\to Y}(Y|Z)\big] \\ &= \sum_{x,y,z} P(x,y,z)\log\frac{P(y|z,x)}{\sum_{x'} P(y|z,x')P(x')}. \end{aligned} \tag{12}$$

Comparing the second expressions in (12) and (10) shows again that the difference between ignoring and cutting is due to the difference between $P(y|z)$ and $\sum_{x'} P(y|z,x')P(x')$.

The following scenario provides a better intuition for the rightmost expression in (12).

EXAMPLE 1 (Redistributing a vaccine). Consider the task of quantifying the effectiveness of a vaccine. Let X indicate whether a patient decides to get vaccinated or not and Y whether the patient becomes infected. Further assume that the vaccine's effectiveness is strongly confounded by age Z because the vaccination often fails for elderly people. At the same time, elderly people request the vaccine more often because they are more afraid of infection. Ignoring other confounders, the DAG in Figure 2(a) visualizes the causal structure.

Deleting the edge $X \to Y$ corresponds to an experiment where the vaccine is randomly assigned to patients regardless of their intent and age (while keeping the total fraction of patients vaccinated constant). Then $P_{X\to Y}(y|z,x) = P_{X\to Y}(y|z) = \sum_{x'} P(y|z,x')P(x')$ represents the conditional probability of infection, given age, *when vaccines are distributed randomly*. $\mathfrak{C}_{X\to Y}$ quantifies the difference to $P(y|z,x)$, which is the conditional probability of infection, given age and intention *when patients act on their intentions*. It thus measures the impact of destroying the coupling between the intention to get the vaccine and getting it via randomized redistribution.

4.2. *Definition via structural equations.* The definition above uses the conditional density $P(x_j|pa_j)$. Estimating a conditional density from empirical data requires huge samples or strong assumptions—particularly for continuous variables.

Fortunately, however, structural equations (also called functional models [14]) allow more direct estimation of causal strength without referring to the conditional distribution.

DEFINITION 3 (Structural equation). A structure equation is a model that explains the joint distribution $P(X_1, \dots, X_n)$ by a deterministic dependence

$$X_j = f_j(PA_j, E_j),$$

where the variables E_j are jointly independent unobserved noise variables. Note that functions f_j that correspond to parentless variables can be chosen to be the identity, that is, $X_j = E_j$.

Suppose that we are given a causal inference method that directly infers the structural equations (e.g., [8, 18]) in the sense that it outputs n-tuples $(e_1^i, \dots, e_n^i)$ with $i = 1, \dots, m$ (with m denoting the sample size) as well as the functions f_j from the observed n-tuples $(x_1^i, \dots, x_n^i)$.

DEFINITION 4 (Removing a causal arrow in a structural equation). Deletion of the arrow $X_k \to X_l$ is modeled by (i) introducing an i.i.d. copy X'_k of X_k and (ii) subsuming the new random variable X'_k into the noise term of f_l. The result is a new set of structural equations:

$$\begin{aligned} x_j &= f_j(pa_j, e_j) \qquad \text{if } j \neq l \text{ and} \\ x_l &= f_l\big(pa_l \setminus \{x_k\}, (x'_k, e_l)\big), \end{aligned} \tag{13}$$

where we have omitted the superscript i to simplify notation.

REMARK 3. To measure the causal influence of a set of arrows, we apply the same procedure after first introducing jointly independent i.i.d. copies of all variables at the tails of deleted arrows.

REMARK 4. The change introduced by the deletion only affects X_l and its descendants, the virtual sample thus keeps all x_j with $j < l$. Moreover, we can ignore all variables X_j with $j > l$ due to Lemma 3.

Note that x'_k must be chosen to be independent of all x_j with $j \leq k$, but, by virtue of the structural equations, not independent of x_l and its descendants. The new structural equations thus generate n-tuples of "virtual" observations $x_1^S, \dots, x_n^S$ from the input

$$\big(e_1, \dots, (x'_k, e_l), \dots, e_n\big).$$

We show below that n-tuples generated this way indeed follow the distribution $P_S(X_1, \dots, X_n)$. We can therefore estimate causal influence via any method that estimates relative entropy using the observed samples $x_1, \dots, x_n$ and the virtual ones $\tilde{x}_1, \dots, \tilde{x}_n$. To illustrate the above scheme, we consider the case where Z and X are causes of Y and we want to delete the edge $X \rightarrow Y$. The case where Y has more than 2 parents follows easily.

EXAMPLE 2 (Two parents). The following table corresponds to the observed variables X, Z, Y, as well as the unobserved noise E^Y which we assumed to be estimated together with learning the structural equations:

$$\left(\begin{array}{cccc} Z & X & E^Y & Y \\ \hline z_1 & x_1 & e_1^Y & f_Y(z_1, x_1, e_1^Y) \\ z_2 & x_2 & e_2^Y & f_Y(z_2, x_2, e_2^Y) \\ \vdots & & & \vdots \\ z_m & x_m & e_m^Y & f_Y(z_m, x_m, e_m^Y) \end{array}\right). \tag{14}$$

To simulate the deletion of $X \rightarrow Y$ we first generate a list of virtual observations for Y after generating samples from an i.i.d. copy X' of X:

$$\left(\begin{array}{ccccc} Z & X & X' & E^Y & Y \\ \hline z_1 & x_1 & x_1' & e_1^Y & f_Y(z_1, x_1', e_1^Y) \\ \vdots & & & & \vdots \\ z_m & x_m & x_m' & e_m^Y & f_Y(z_m, x_m', e_m^Y) \end{array}\right). \tag{15}$$

A simple method to simulate the i.i.d. copy is to apply some random permutation $\pi \in S_m$ to $x_1, \dots, x_n$ and obtain $x_{\pi(1)}, \dots, x_{\pi(n)}$ (see [9], S.1). Deleting several arrows with source node X requires several identical copies $X', X'', \dots$ of X, each generated by a different permutation.

We then throw away the two noise columns, that is, the original noise E^Y and the additional noise X':

$$\left(\begin{array}{ccc} Z & X & Y \\ \hline z_1 & x_1 & f_Y(z_1, x_1', e_1^Y) \\ \vdots & & \vdots \\ z_m & x_m & f_Y(z_m, x_m', e_m^Y) \end{array}\right). \tag{16}$$

To see that this triple is indeed sampled from the desired distribution $P_S(X, Y, Z)$, we recall that the original structural equation simulates the conditional $P(Y|X, Z)$. After inserting X' we obtain the new conditional $\sum_{x'} P(Y|x', Z) \times P(x')$. Multiplying it with $P(X, Z)$ yields $P_S(X, Y, Z)$, by definition. Using the above samples from $P_S(X, Y, Z)$ and samples from $P(X, Y, Z)$ we can estimate

$$\mathfrak{C}_{X \rightarrow Y} = D\bigl(P(X, Y, Z) \| P_S(X, Y, Z)\bigr)$$

using some known schemes [16] for estimating relative entropies from empirical data. It is important that the samples from the two distributions are disjoint, meaning that we need to split the original sample into two halves, one for P and one for P_S.

The generation of P_S for a set S of arrows works similarly: every input of a structural equation that corresponds to an arrow to be removed is fed with an independent copy of the respective variable. Although it is conceptually simple to estimate causal strength by generating the entire joint distribution P_S, Theorem 5(a) will show how to break the problem into parts that make estimation of relative entropies from finite data more feasible.

We now revisit mediation analysis [2, 14, 19], which is also based on structural equations, and mention an interesting relation to our work. Although we have pointed out that intervening by "cutting edges" is complementary to the intervention on nodes considered there, distributions like P_S can also occur in an implicit way. To explore the indirect effect $X \rightarrow Z \rightarrow Y$ in Figure 2(b), one can study the effect of X on Y in the reduced DAG $X \rightarrow Z \rightarrow Y$ under the distribution $P_{X\rightarrow Y}$ or under the distribution obtained by setting the copy X' to some fixed value x'. Remarkably, cutting $X \rightarrow Y$ is then used to study the strength of the *other path* while we use it to study the strength[2] of $X \rightarrow Y$.

4.3. *Properties of causal strength.* This subsection shows that our definition of causal strength satisfies postulates P0–P4. We observe at the same time some other useful properties. We start with a property that is used to show P0.

Causal strength majorizes observed dependence. Recalling that $P(X_1, \dots, X_n)$ factorizes into $\prod_j P(X_j|PA_j)$ with respect to the true causal DAG G, one may ask how much error one would cause if one was not aware of all causal influences and erroneously assumed that the true DAG would be the one where some set S of arrows is missing. The conditionals with respect to the reduced set of parents define a different joint distribution.

DEFINITION 5 (Distribution after ignoring arrows). Given distribution P Markovian with respect to G and set of arrows S, let the partially observed distribution (where interactions across S are hidden) for node X_j be

$$\tilde{P}_S(x_j|pa_j^{\bar{S}}) = \sum_{pa_j^S} P(x_j|pa_j^S, pa_j^{\bar{S}}) P(pa_j^S|pa_j^{\bar{S}}).$$

Let the partially observed distribution for all the nodes be the product

$$\tilde{P}_S(x_1, \dots, x_n) = \prod_j \tilde{P}_S(x_j|pa_j^{\bar{S}}). \tag{17}$$

[2] We are grateful to an anonymous referee for this observation.

REMARK 5. Intuitively, the observed influence of a set of arrows should be quantified by comparing the data available to an observer who can see the entire DAG with the data available to an observer who sees all the nodes of the graph, but only some of the arrows. Definition 5 formalizes "seeing only some of the arrows."

Building on Remark 2, the definition of the *observed dependence* of a set of arrows takes the same general form as for causal influence. However, instead of inserting noise on the arrows, we instead simply prevent ourselves from seeing them.

DEFINITION 6 (Observed influence). Given a distribution P that is Markovian with respect to G and set of arrows S, let the observed influence of the arrows in S be

$$\mathfrak{O}_S(P) := D(P \| \tilde{P}_S),$$

with $\tilde{P}_S$ defined in (17).

The following result, proved in Section A.1, is crucial to proving P0.

THEOREM 2 (Causal influence majorizes observed dependence). *Causal influence decomposes into observed influence plus a* nonnegative *term quantifying the divergence between the partially observed and interventional distributions*

$$\mathfrak{C}_S(P) = \mathfrak{O}_S(P) + \sum_{j=1}^{n} P\big(pa_j^{\bar{S}}\big) \cdot D\big(\tilde{P}_S\big(X_j | pa_j^{\bar{S}}\big) \| P_S\big(X_j | pa_j^{\bar{S}}\big)\big). \tag{18}$$

The theorem shows that "snapping upstream dependencies" by using purely local data that is, by marginalizing using the distribution of the source node $P(X_i)$ rather than the conditional $P(X_i | PA_i)$—is essential to quantifying causal influence.

Proof of postulates for causal strength.

P0: Let G_S be the DAG obtained by removing the arrows in S from G. Let $PA_j^{\bar{S}}$ be the parents of X_j in G_S, that is, those that are not in S and introduce the set of nodes Z_j such that $PA_j = PA_j^{\bar{S}} \cup Z_j$. By Theorem 2, $\mathfrak{C}_S = 0$ implies $\mathfrak{O}_S = 0$, that is, $\tilde{P}_S = P$, which implies

$$P(X_j | pa_j) = P\big(X_j | pa_j^{\bar{S}}\big) \qquad \forall pa_j^{\bar{S}} \text{ with } P\big(pa_j^{\bar{S}}\big) \neq 0, \tag{19}$$

that is, $X_j \perp\!\!\!\perp Z_j | PA_j^{\bar{S}}$.

We use again the Ordered Markov condition

$$X_j \perp\!\!\!\perp PR_j | PA_j \qquad \forall j, \tag{20}$$

where PR_j denote the predecessors of X_j with respect to some ordering of nodes that is consistent with G. By the contraction rule [14], (19) and (20) yields

$$X_j \perp\!\!\!\perp PR_j \cup Z_j \,|\, PA_j^{\bar{S}},$$

and hence

$$X_j \perp\!\!\!\perp PR_j \,|\, PA_j^{\bar{S}},$$

which is the Ordered Markov condition for G_S if we use the same ordering of nodes for G_S.

P1: One easily checks $\mathfrak{C}_{X\to Y} = I(X;Y)$ for the 2-node DAG $X \to Y$, because $P_{X\to Y}(x,y) = P(x)P(y)$, and thus

$$D(P\|P_{X\to Y}) = D\bigl(P(X,Y)\|P(X)P(Y)\bigr) = I(X;Y).$$

P2: Follows from the following lemma.

LEMMA 3 (Causal strength as local relative entropy). *Causal strength $\mathfrak{C}_{k\to l}$ can be written as the following relative entropy distance or conditional relative entropy distance*:

$$\begin{aligned}\mathfrak{C}_{k\to l} &= \sum_{pa_l} D\bigl[P(X_l|pa_l)\|P_S(X_l|pa_l)\bigr]P(pa_l)\\ &= D\bigl[P(X_l|PA_l)\|P_S(X_l|PA_l)\bigr].\end{aligned}$$

Note that $P_S(X_l|pa_l)$ actually depends on the reduced set of parents $PA_l \setminus X_k$ only, but it is more convenient for the notation and the proof to keep the formal dependence on all PA_l.

PROOF OF LEMMA 3.

$$\begin{aligned}D(P\|P_S) &= \sum_{x_1\cdots x_n} P(x_1\cdots x_n)\log\frac{P(x_1\cdots x_n)}{P_S(x_1\cdots x_n)}\\ &= \sum_{x_1\cdots x_n} P(x_1\cdots x_n)\log\prod_{j=1}^{n}\frac{P(x_j|pa_j)}{P_S(x_j|pa_j)}\\ &= \sum_{j=1}^{n}\sum_{x_j,pa_j} P(x_j,pa_j)\log\frac{P(x_j|pa_j)}{P_S(x_j|pa_j)}\\ &= \sum_{j=1}^{n} D\bigl[P(X_j|PA_j)\|P_S(X_j|PA_j)\bigr].\end{aligned}$$

For all $j \neq l$ we have $D[P(X_j|PA_j)\|P_S(X_j|PA_j)] = 0$, because $P(X_l|PA_l)$ is the only conditional that is modified by the deletion. □

P3: Apart from demonstrating the postulated inequality, the following result shows that we have the equality $\mathfrak{C}_{X\to Y} = I(X;Y|PA_Y^X)$ for independent causes. To keep notation simple, we have restricted our attention to the case where Y has only two causes X and Z, but Z can also be interpreted as representing all parents of Y other than X.

THEOREM 4 (Decomposition of causal strength). *For the DAGs in Figure 2, we have*

$$\mathfrak{C}_{X\to Y} = I(X;Y|Z) + D\big[P(Y|Z)\|P_{X\to Y}(Y|Z)\big]. \tag{21}$$

If X and Z are independent, the second term vanishes.

PROOF. Equation (21) follows from Theorem 2: First, we observe $\mathfrak{D}_S(P) = I(X;Y|Z)$ because both measure the relative entropy distance between $P(X,Y,Z)$ and $\tilde{P}_S(X,Y,Z) = P(X,Z)P(Y|Z)$. Second, we have

$$P_S(X,Y,Z) = P(X,Z)P_{X\to Y}(Y|Z).$$

The second summand in (18) reduces to

$$\begin{aligned}&\sum_z P(z)D\big[\tilde{P}_S(Y|z)\|P_S(Y|z)\big]\\ &\quad= \sum_z P(z)D\big[P(Y|z)\|P_S(Y|z)\big]\\ &\quad= D\big[P(Y|Z)\|P_S(Y|Z)\big].\end{aligned}$$

To see that the second term in equation (21) vanishes for independent X, Z, we observe

$P_{X\to Y}(Y|Z) = P(Y|Z)$ because

$$P_{X\to Y}(y|z) = \sum_x P(y|x,z)P(x) = \sum_x P(y|x,z)P(x|z) = P(y|z). \qquad \square$$

Theorem 4 states that conditional mutual information underestimates causal strength. Assume, for instance, that X and Z are almost always equal because Z has such a strong influence on X that it is an almost perfect copy of it. Then $I(X;Y|Z) \approx 0$ because knowing Z leaves almost no uncertainty about X. In other words, strong dependencies between the causes X and Z makes the influence of cause X almost invisible when looking at the conditional mutual information $I(X;Y|Z)$ only. The second term in (21) corrects for the underestimation. When X depends deterministically on Z, it is even the only remaining term (here, we have again assumed that the conditional distributions are defined for events that do not occur in observational data).

To provide a further interpretation of Theorem 4, we recall that $I(X;Y|Z)$ can be seen as the impact of ignoring the edge $X \to Y$; see Remark 2. Then the impact of cutting $X \to Y$ is given by the impact of ignoring this link plus the impact that cutting has on the conditional $P(Y|Z)$.

P4: This postulate is part (d) of the following collection of results that relates strength of sets to its subsets.

THEOREM 5 (Relation between strength of sets and subsets). *The causal influence given in Definition 2 has the following properties*:

(a) Additivity regarding targets. *Given set of arrows S, let $S_i = \{s \in S | \operatorname{trg}(s) = X_i\}$, then*

$$\mathfrak{C}_S(P) = \sum_i \mathfrak{C}_{S_i}(P).$$

(b) Locality. *Every $\mathfrak{C}_{S_i}$ only depends on the conditional $P(X_i|PA_i)$ and the joint distribution of all parents $P(PA_i)$.*

(c) Monotonicity. *Given sets of arrows $S_1 \subset S_2$ targeting single node Z, such that the source nodes in S_1 are jointly independent and independent of the other parents of Z. Then we have*

$$\mathfrak{C}_{S_1}(P) \leq \mathfrak{C}_{S_2}(P).$$

(d) Heredity property. *Given sets of arrows $S \subset T$, we have*

$$\mathfrak{C}_T(P) = 0 \quad \Longrightarrow \quad \mathfrak{C}_S(P) = 0.$$

The proof is presented in Appendix A.3. The intuitive meaning of these properties is as follows. Part (a) says that causal influence is additive if the arrows have different targets. Otherwise, we can still decompose the set S into equivalence classes of arrows having the same target and obtain additivity regarding the decomposition. This can be helpful for practical applications because estimating each $D[P(PA_i, X_i) \| P_{S_i}(PA_i, X_i)]$ from empirical data requires less data then estimating the distance $D(P \| P_S)$ for the entire high dimensional distributions.

We will show in Section 4.4 that general additivity fails. Part (b) is an analog of P2 for multiple arrows. According to (c), the strength of a subset of arrows cannot be smaller than the strength of its superset, provided that there are no dependencies among the parent nodes. Finally, part (d) is exactly our postulate P4.

Parts (c) and (d) suggest that monotonicity may generalize to the case of dependent parents: $S \subset T \Longrightarrow \mathfrak{C}_S(P) \leq \mathfrak{C}_T(P)$. However, the following counterexample due to Bastian Steudel shows this is not the case.

EXAMPLE 3 (XOR—counterexample to monotonicity when parents are dependent). Consider the DAG(a) in Figure 2 and let the relation between X, Y, Z be given by the structural equations

$$X = Z, \tag{22}$$

$$Y = X \oplus Z. \tag{23}$$

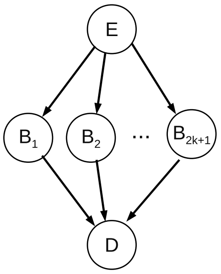

FIG. 6. *Causal structure of an error-correcting scheme*: *the encoder generates* $2k+1$ *bits from a single one*. *The decoder decodes the* $2k+1$ *bit words into a single bit again*.

Let $P(Z=0)=a$ and $P(Z=1)=1-a$. Letting $S=\{Z\to X\}$ and $T=\{Z\to X, X\to Y\}$ we find that

$$\mathfrak{C}_S(P) = -a\log(a) - (1-a)\log(1-a) \quad \text{and}$$
$$\mathfrak{C}_T(P) = -\log\bigl(a^2+(1-a)^2\bigr).$$

For $a \notin \{\frac{1}{2}, 0, 1\}$, strict concavity of the logarithm implies $\mathfrak{C}_T(P) < \mathfrak{C}_S(P)$.

4.4. *Examples and paradoxes*. *Failure of subadditivity*: The strength of a set of arrows is not bounded from above by the sum of strength of the single arrows. It can even happen that removing one arrow from a set has no impact on the joint distribution while removing all of them has significant impact, which occurs in communication scenarios that use redundancy.

EXAMPLE 4 (Error correcting code). Let E and D be binary variables that we call "encoder" and "decoder" (see Figure 6) communicating over a channel that consists of the bits $B_1, \dots, B_{2k+1}$. Using the simple repetition code, all B_j are just copies of E. Then D is set to the logical value that is attained by the majority of B_j. This way, k errors can be corrected, that is, removing k or less of the links $B_j \to D$ has no effect on the joint distribution, that is, $P_S = P$ for $S := (B_1 \to D, B_2 \to D, \dots, B_k \to D)$, hence $\mathfrak{C}_S(P) = 0$. In words: removing k or less arrows is without impact, but removing all of them is, of course. After all, the arrows jointly generate the dependence $I(E; D) = I((E, B_1, \dots, B_k); D) = 1$, provided that E is uniformly distributed.

Clearly, the outputs of E causally influence the behavior of D. We therefore need to consider interventions that destroy many arrows at once if we want to capture the fact that their joint influence is nonzero.

Thus, causal influence of arrows is *not* subadditive: the strength of each arrow $B_j \to D$ is zero, but the strength of the set of all $B_j \to D$ is 1 bit.

Failure of superadditivity: The following example reveals an opposing phenomenon, where the causal strength of a set is smaller than the sum of the single arrows.

EXAMPLE 5 (XOR with uniform input). Consider the structural equations (22) and (23) with uniformly distributed Z. The causal influence of each arrow targeting the XOR-gate individually is the same as the causal influence of both arrows taken together:

$$\mathfrak{C}_{X\to Y}(P)=\mathfrak{C}_{Z\to Y}(P)=\mathfrak{C}_{\{X\to Y,Z\to Y\}}(P)=1 \text{ bit}.$$

Strong influence without dependence/failure of converse of P0: Revisiting Example 5 is also instructive because it demonstrates an extreme case of confounding where $I(X;Y|Z)$ vanishes but causal influence is strong. Removing $X\to Y$ yields

$$P_{X\to Y}(x,y,z)=P(x,z)P(y),$$

where $P(z)=P(y)=1/2$ and $P(x|z)=\delta_{x,z}$. It is easy to see that

$$D(P\|P_{X\to Y})=1,$$

because P is a uniform distribution over 2 possible triples (x,y,z), whereas $P_{X\to Y}$ is a uniform distribution over a superset of 4 triples.

The impact of cutting the edge $X\to Y$ is remarkable: both distributions, the observed one P as well as the post-cutting distribution P_S, factorize $P_S(X,Y,Z)=P_S(X,Z)P_S(Y)$ and $P(X,Y,Z)=P(X,Z)P(Y)$. Cutting the edge keeps this product structure and changes the joint distributions by only changing the marginal distribution of Y from $P(Y)$ to $P_S(Y)$.

Note that P satisfies the Markov condition with respect to $G_{X\to Y}$ (i.e., the DAG obtained from the original one by dropping $X\to Y$) because Y is a constant. Since $\mathfrak{C}_{X\to Y}\neq 0$, this shows that the converse of P0 does not hold.

Strong effect of little information: The following example considers multiple arrows and shows that their joint strength may even be strong when they carry the same small amount of information.

EXAMPLE 6 (Broadcasting). Consider a single source X with many targets $Y_1,\dots,Y_n$ such that each Y_i copies X, see Figure 7. Assume $P(X=0)=P(X=1)=\frac{1}{2}$. If S is the set of all arrows $X\to Y_j$ then $\mathfrak{C}_S=n$. Thus, the single node X exerts n bits of causal influence on its dependents.

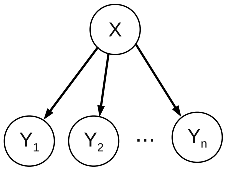

FIG. 7. *Broadcasting one bit from one node to multiple nodes.*

5. Causal influence between two time series.

5.1. *Definition.* Since causal analysis of time series is of high practical importance, we devote a section to this case. For some fixed t, we introduce the short notation $X \to Y_t$ for the set of all arrows that point to Y_t from some X_s with $s < t$. Then

$$\mathfrak{C}_{X \to Y_t}$$

measures the impact of deleting all these arrows. We propose to replace transfer entropy with this measure since it does not suffer from the drawbacks described in Section 3.3.

Section 4.2 describes how to estimate causal strength from finite data for one arrow and briefly mentions how this generalizes to set of arrows. To keep this section self-consistent, we briefly rephrase the description for the case of time series.

Suppose we have learned the structural equation model

$$Y_t = f_t(X_{t-1}, X_{t-2}, \dots, X_{t-p}, E_t), \tag{24}$$

from observed data $(x_t, y_t)_{t \leq 0}$, where the noise variables E_t are jointly independent and independent of $X_t, X_{t-1}, \dots, Y_{t-1}, Y_{t-2}, \dots$. Assume, moreover, that we have inferred the corresponding values $(e_t)_{t \leq 0}$ of the noise. If we have multiple copies of the time series, we can apply the method described in Section 4.2 in a straightforward way: Due to the locality property stated in Theorem 5(b), we only consider the variables $X_{t-p}, \dots, X_{t-1}, Y_t$ and feed (24) with i.i.d. copies of $X_{t-p}, \dots, X_{t-1}$ by applying random permutations to the observations, which then yields samples from the modified distribution $P_S(X_{t-p}, \dots, X_{t-1}, Y_t)$.

If we have only one observation for each time instance, we have to assume stationarity (with constant function $f_t = f$) and ergodicity and generate an artificial statistical sample by looking at sufficiently distant windows.

5.2. *Comparison of causal influence with transfer entropy.* We first recall the example given by [4] showing a problem with transfer entropy (Section 3.3). Assume that the variables X_t, Y_t in Figure 3, right, are binary and the transition from X_{t-1} to Y_t is a perfect copy and likewise the transition from Y_{t-1} to X_t. Assume, moreover, that the two causal chains have been initialized such that, with probability 1/2, all variables are 1 and with probability 1/2 all are zero. Then the set $X \to Y_t$ is the singleton $S := \{X_{t-1} \to Y_t\}$. Using Lemma 3, we have

$$\mathfrak{C}_{X_{t-1} \to Y_t} = D\big[P(Y_t, X_{t-1}) \| P_S(Y_t, X_{t-1})\big].$$

Since Y_t is a perfect copy of X_{t-1}, we have

$$P(y_t, x_{t-1}) = \begin{cases} 1/2, & \text{for } x_{t-1} = y_t, \\ 0, & \text{otherwise} \end{cases}$$

into

$$P_S(y_t, x_{t-1}) = 1/4 \qquad \text{for } (y_t, x_{t-1}) \in \{0, 1\}^2.$$

One easily checks $D(P \| P_S) = 1$.

Note that the example is somewhat unfair, since it is *impossible* to distinguish the structural equations from a model without interaction between X and Y, where X_{t+1} is obtained from X_t by inversion and similarly for Y, no matter how many observations are performed. Thus, from observing the system it is impossible to tell whether or not X exerts an influence on Y. However, the following modification shows that transfer entropy still goes quantitatively wrong if small errors are introduced.

EXAMPLE 7 (Perturbed transfer entropy counterexample). Perturb Ay and Polani's example by having Y_t copy X_{t-1} correctly with probability $p = 1 - \varepsilon$. Set node Y_t's transitions as Markov matrix

$$\left(\begin{array}{c|cc} & x_{t-1}=0 & x_{t-1}=1 \\ \hline y_t=0 & 1-\varepsilon & \varepsilon \\ y_t=1 & \varepsilon & 1-\varepsilon \end{array}\right),$$

and similarly for the transition from Y_{t-1} to X_t.

The transfer entropy from X to Y at time t is

$$\begin{aligned} \mathrm{TE} &= I(X_{(-\infty,t-1]}; Y_t | Y_{(-\infty,t-1]}) = I(X_{t-1}; Y_t | Y_{t-2}) \\ &= H(Y_t|Y_{t-2}) - H(Y_t|Y_{t-2}, X_{t-1}) = H(Y_t|Y_{t-2}) - H(Y_t|X_{t-1}), \end{aligned}$$

where $H(\cdot|\cdot)$ denotes the conditional Shannon entropy. The equalities can be derived from d-separation in the causal DAG Figure 3, right [14]. For instance, conditioning on Y_{t-2}, renders the pair (Y_t, X_{t-1}) independent of all the remaining past of X and Y. We find

$$\begin{aligned} -H(Y_t|X_{t-1}) &= \varepsilon \log \varepsilon + (1-\varepsilon)\log(1-\varepsilon), \\ H(Y_t|Y_{t-2}) &= 2\varepsilon(1-\varepsilon)\log\frac{1}{2\varepsilon(1-\varepsilon)} + \left(1 - 2\varepsilon + 2\varepsilon^2\right)\log\frac{1}{1-2\varepsilon+2\varepsilon^2}. \end{aligned}$$

Hence,

$$\begin{aligned} \mathrm{TE} = {} & \left(1 - 2\varepsilon + 2\varepsilon^2\right)\log\frac{1}{1-2\varepsilon+2\varepsilon^2} + 2\varepsilon(1-\varepsilon)\log\frac{1}{2\varepsilon(1-\varepsilon)} \\ & + \varepsilon\log\varepsilon + (1-\varepsilon)\log(1-\varepsilon), \end{aligned}$$

which tends to zero as $\varepsilon \to 0$.

Causal influence, on the other hand, is given by the mutual information $I(Y_t; X_{t-1})$ because all edges other than $X_{t-1} \to Y_t$ are irrelevant (see Postulate P2). Thus,

$$\mathfrak{C}_{X\to Y_t} = H(Y_t) - H(Y_t|X_{t-1}) = 1 + (1-\varepsilon)\log(1-\varepsilon) + \varepsilon\log\varepsilon,$$

which tends to 1 for $\varepsilon \to 0$. Hence, causal influence detects the causal interactions between X and Y based on *empirical data*, whereas transfer entropy does not. Thanks to the perturbation, the joint distribution tells us the kind of causal relations by which it is generated. For large enough samples, the strong discrepancy between transfer entropy and our causal strength thus becomes apparent.

6. Causal strength for linear structural equations. For linear structural equations, we can provide a more explicit expression of causal strength under the assumption of multivariate Gaussianity. Let n random variables $X_1, \ldots, X_n$ be ordered such that there are only arrows from X_i to X_j for $i < j$. Then we have structural equations

$$X_j = \sum_{i<j} A_{ij} X_i + E_j,$$

where all E_j are jointly independent noise variables. In vector and matrix notation we have

$$X = AX + E \qquad \text{that is, } X = (I - A)^{-1} E, \tag{25}$$

where A is lower triangular with zeros in the diagonal.

To compute the strength of $S \subset \{1, \ldots, n\}$, we assume for reasons of convenience that all variables have zero mean. Then $D(P \| P_S)$ can be computed from the covariance matrices alone.

The covariance matrix of X reads

$$\Sigma = (I - A)^{-1} \Sigma_E (I - A)^{-T},$$

where Σ_E denotes the covariance matrix of the noise (which is diagonal by assumption) and $(\cdot)^{-T}$ the transpose of the inverse of a matrix.

To compute the covariance matrix Σ^S of P_S, we first split A into $A_S + A_{\bar{S}}$, where A_S contains only those entries that correspond to the edges in the set S and $A_{\bar{S}}$ only those corresponding to the complement of S. Using this notation, the modified structural equations read

$$X = A_{\bar{S}} X + E + A_S X', \tag{26}$$

where $X' = (X'_1, \ldots, X'_n)^T$ and each X'_j has the same distribution as X_j and satisfies joint independence of all $X'_1, \ldots, X'_n, E_1, \ldots, E_n$. It is convenient to define the modified noise

$$E' := E + A_S X',$$

with covariance matrix

$$\Sigma_{E'} = \Sigma_E + A_S \Sigma_X^D A_S^T, \tag{27}$$

where Σ_X^D contains only the diagonal entries of Σ_X (recall that all X'_j are independent). The modified variables X^S are now given by the equation

$$X^S = A_{\bar{S}} X + E',$$

which formally looks like (25), although the components of E' are dependent while the E_j in (25) are independent. Thus, we obtain the modified covariance matrix of X by

$$\Sigma_S = (I - A_{\bar{S}})^{-1} \Sigma_{E'} (I - A_{\bar{S}})^{-T}.$$

The causal strength now reads

$$\begin{aligned}\mathfrak{C}_S = D(P \| P_S) &= \frac{1}{2}\left(\operatorname{tr}[\Sigma_S^{-1}\Sigma] - \log\frac{\det\Sigma}{\det\Sigma_S} - n\right)\\ &= \frac{1}{2}\biggl(\operatorname{tr}\bigl[(I - A_{\bar{S}})\Sigma_{E'}^{-1}(I - A_{\bar{S}})(I - A)^{-1}\Sigma_E(I - A)^{-1}\bigr]\\ &\qquad - \log\frac{\det(I - A)^{-1}\Sigma_E(I - A)^{-1}}{\det(I - A_{\bar{S}})^{-1}\Sigma_{E'}(I - A_{\bar{S}})^{-1}} - n\biggr),\end{aligned}$$

with $\Sigma_{E'}$ given by (27).

EXAMPLE 8 (Linear structural equations with independent parents). It is instructive to look at the following simple case:

$$X_n := \sum_j \alpha_{nj} X_j + E_n \qquad \text{with } E_n, X_1, \ldots, X_{n-1} \text{ jointly independent.}$$

For the set $S := \{X_1 \to X_n, \ldots, X_k \to X_n\}$ with $k \leq n$ some calculations show

$$\mathfrak{C}_S = \frac{1}{2}\log\frac{\operatorname{Var}(X_n) - \sum_{j=k+1}^{n-1}\alpha_{nj}^2\operatorname{Var}(X_j)}{\operatorname{Var}(X_n) - \sum_{j=1}^{n-1}\alpha_{nj}^2\operatorname{Var}(X_j)}.$$

For the single arrow $X_1 \to X_n$, we thus obtain

$$\mathfrak{C}_{X_1 \to X_n} = \frac{1}{2}\log\frac{\operatorname{Var}(X_n) - \sum_{j=2}^{n-1}\alpha_{nj}^2\operatorname{Var}(X_j)}{\operatorname{Var}(X_n) - \sum_{j=1}^{n-1}\alpha_{nj}^2\operatorname{Var}(X_j)}.$$

If X_1 is the only parent, that is, $n = 2$, we have

$$\mathfrak{C}_{X_1 \to X_2} = \frac{1}{2}\log\frac{\operatorname{Var}(X_2)}{\operatorname{Var}(X_2) - \alpha_{21}^2\operatorname{Var}(X_1)} = -\frac{1}{2}\log(1 - r_{21}),$$

with r_{21} as in equation (3) introduced in the context of ANOVA. Note that the relation between our measure and r_{n1} is less simple for $n > 2$ because r_{n1} would then still measure the fraction of the variance of X_n explained by X_1, while $\mathfrak{C}_{X_1 \to X_n}$ is related to the fraction of the *conditional* variance of X_n, given its other parents, explained by X_1. This is because our causal strength reduces to a *conditional* mutual information for independent parents; see the last sentence of Theorem 4.

7. Experiments. Code for all experiments can be downloaded at http://webdav.tuebingen.mpg.de/causality/.

7.1. *DAGs without time structure.* We here restrict attention to linear structural equations, but interesting generalizations are given by additive noise models [8, 17, 18] and post-nonlinear models [23].

The first step in estimating the causal strength consists in inferring the structure matrix A in (25) from the given matrix $\mathbf{X}$ of observations x_j^i with $j = 1, \dots, n$ and $i = 1, \dots, 2k$ (the jth row corresponds to the observed values of X_j). We did this step by ridge regression. We decompose A into the sum $A_S + A_{\bar{S}}$ as in Section 6.

Then we divide the columns of $\mathbf{X}$ into two parts $\mathbf{X}_A$ and $\mathbf{X}_B$ of sample size k. While $\mathbf{X}_A$ is kept as it is, $\mathbf{X}_B$ is used to generate new samples according to the modified structural equations: First, we note that the values of the noise variables corresponding to the observations $\mathbf{X}_B$ are given by the residuals

$$\mathbf{E}_B := \mathbf{X}_B - A \cdot \mathbf{X}_B.$$

Then we generate a matrix $\mathbf{X'_B}$ by applying independent random permutations to the columns of $\mathbf{X}_B$, which simulates samples of the random variables X'_j in (26). Samples from the modified structural equation are now given by

$$\mathbf{X}_B^S := (I - A_{\bar{S}})^{-1} \cdot \mathbf{X}_B + \mathbf{E}_B + A_S \cdot \mathbf{X}'_B.$$

To estimate the relative entropy distance between P and P_S (with samples $\mathbf{X}_A$ and $\mathbf{X}_B^S$), we use the method described in [16]: Let d_i be the euclidean distance from the ith column in $\mathbf{X}_A$ to the rth nearest neighbor among the other columns of $\mathbf{X}_A$ and d_i^S be the distance to the rth nearest neighbor among all columns of $\mathbf{X}_B$, then the estimator reads

$$\hat{D}(P \| P_S) := \frac{n}{k} \sum_{i=1}^{k} \log \frac{d_i^S}{d_i} + \log \frac{k}{k-1}.$$

Figure 8 shows the difference between estimated and computed causal strength for the simplest DAG $X_1 \to X_2$ with increasing structure coefficient. For some edges, we obtain significant bias. However, since the bias depends on the distributions [16], it would be challenging to correct for it.

To provide a more general impression on the estimation error, we have considered a complete DAG on $n = 3$ and $n = 6$ nodes and randomly generated structure coefficients. In each of $\ell = 1, \dots, 100$ runs, the structure matrix is generated by independently drawing each entry from a standard normal distribution. For each of the $\binom{n}{2}$ arrows $i \to j$ and each ℓ we computed and estimated $\mathfrak{C}_{i \to j}$, which yields the x-value and the y-value, respectively, of one of the $100 \cdot \binom{n}{2}$ points in the scatter plots in Figure 9. Remarkably, we do not see a significant degradation for $n = 6$ nodes (right) compared to $n = 3$ (left).

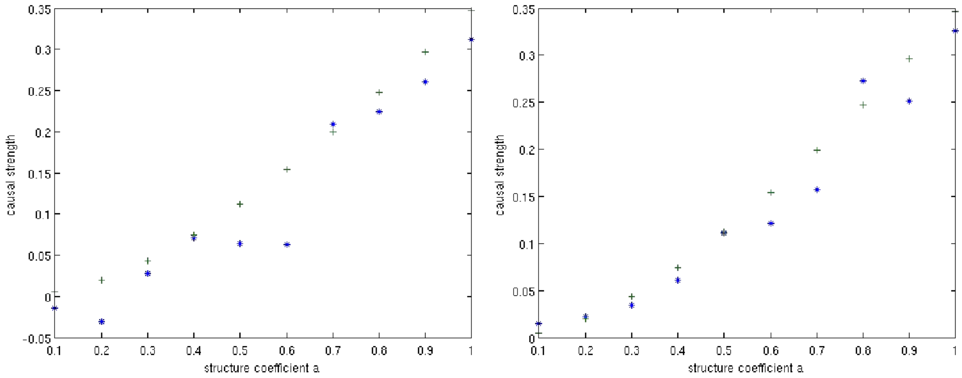

FIG. 8. *Estimated and computed value* $\mathfrak{C}_{1\to 2}$ *for* $X_1 \to X_2$, *indicated by* $*$ *and* $+$, *respectively. The underlying linear Gaussian model reads* $X_2 = a \cdot X_1 + E$. *Left for sample size* 1000, *which amounts to* 500 *samples in each part. Right*: *sample size* 2000, *which yields more reliable results.*

7.2. *Time series.* The fact that transfer entropy fails to capture causal strength has been one of our motivations for defining a different measure. We revisit the critical example in Section 5.2, where the dynamical evolution on two bits was given by noisy copy operations from X_{t-1} to Y_t and Y_{t-1} to X_t. This way, we obtained causal strength 1 bit when the copy operations is getting perfect. Our software for estimating causal strength only covers the case of linear structural equations, with the additional assumption of Gaussianity for the subroutines that *compute* the causal strength from the covariance matrices for comparison with the estimated value.

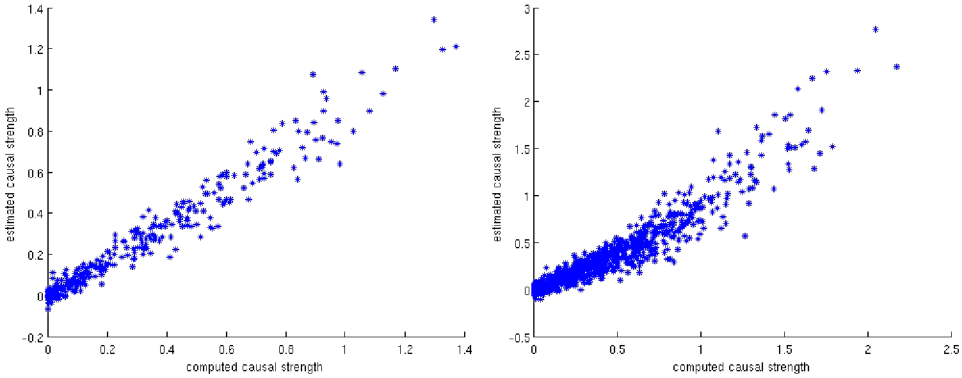

FIG. 9. *Relation between computed and estimated single arrow strengths for* 100 *randomly generated structure matrices and noise variance* 1. *The estimation is based on sample size* 1000. *Left*: *complete DAG on* 3 *nodes. Right*: *the same for* 5 *nodes.*

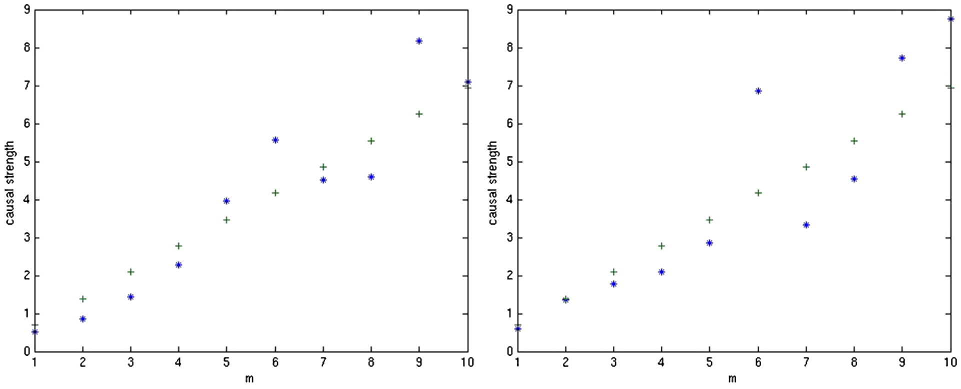

FIG. 10. *Estimated and computed value* $\mathfrak{C}_{X\to Y_t}$ *where* $\varepsilon = 2^{-m}$ *and* m *runs from* 1 *to* 10. *Left*: *for length* $T = 5000$. *Right*: $T = 50{,}000$.

A natural linear version of Example 7 is an autoregressive (AR-) model of order 1 given by

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} 0 & \sqrt{1-\varepsilon^2} \\ \sqrt{1-\varepsilon^2} & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} E_t^X \\ E_t^Y \end{pmatrix},$$

where E_t^X, E_t^Y are independent noise terms. We consider the stationary regime where X_t and Y_t have unit variance and E_t has variance ε^2. For $\varepsilon \to 0$ the influence from X_{t-1} on Y_t, and similarly from Y_{t-1} to X_t gets deterministic. We thus obtain infinite causal strength (note that two deterministically coupled random variables with probability density have infinite mutual information). It is easy to see that transfer entropy does not diverge, because the conditional variance of Y_t is $2\varepsilon^2$ if only the past of Y is given and ε^2 if the past of X is given in addition. Reducing the variance by the factor 2 corresponds to the constant information gain of $\frac{1}{2}\log 2$, regardless of how small ε is.

Figure 10 shows the computed and estimated values of causal strength for decreasing ε, that is, the deterministic limit. Note that, in this limit, the estimated relative entropy can deviate strongly from the true one because the true one diverges since P_S lives on a higher dimensional manifold than P. This probably explains the large errors for $m \geq 6$, which correspond to quite low noise level already.

8. Conclusions.

We have defined the strength of an arrow or a set of arrows in a causal Bayesian network by quantifying the impact of an operation that we called "destruction of edges". We have stated a few postulates that we consider natural for a measure of causal strength and shown that they are satisfied by our measure. We do not claim that our list is complete, nor do we claim that measures violating our postulates are inappropriate. How to quantify causal influence may strongly depend on the purpose of the respective measure.

For a brief discussion of an alternative measure of causal strength and some of the difficulties that arising when quantifying the total influence of one set of nodes on another, see the supplementary material [9].

The goal of this paper is to encourage discussions on how to define causal strength within a framework that is general enough to include dependencies between variables of arbitrary domains, including nonlinear interactions, and multidimensional and discrete variables at the same time.

APPENDIX: FURTHER PROPERTIES OF CAUSAL STRENGTH AND PROOFS

A.1. Proof of Theorem 2. Expand $\mathfrak{C}_S(P)$ as

$$D(P\|P_S) = \sum_{x_1\cdots x_n} P(x_1\cdots x_n)\log\frac{P(x_1\cdots x_n)}{P_S(x_1\cdots x_n)} \tag{28}$$

$$\begin{aligned} &= \sum_{x_1\cdots x_n} P(x_1\cdots x_n)\log\frac{P(x_1\cdots x_n)}{\tilde{P}_S(x_1\cdots x_n)} \\ &\quad + \sum_{x_1\cdots x_n} P(x_1\cdots x_n)\log\frac{\tilde{P}_S(x_1\cdots x_n)}{P_S(x_1\cdots x_n)}. \end{aligned} \tag{29}$$

Note that the second term can be written as

$$\begin{aligned} &\sum_{x_1\cdots x_n} P(x_1\cdots x_n)\log\prod_{j=1}^{n}\frac{\tilde{P}_S(x_j|pa_j^{\bar{S}})}{P_S(x_j|pa_j^{\bar{S}})} \\ &\quad = \sum_{j=1}^{n}\sum_{x_1\cdots x_n} P(x_1\cdots x_n)\log\frac{\tilde{P}_S(x_j|pa_j^{\bar{S}})}{P_S(x_j|pa_j^{\bar{S}})} \end{aligned} \tag{30}$$

$$= \sum_{j=1}^{n}\sum_{x_j,pa_j} P(x_j, pa_j^{\bar{S}}, pa_j^{S})\log\frac{\tilde{P}_S(x_j|pa_j^{\bar{S}})}{P_S(x_j|pa_j^{\bar{S}})} \tag{31}$$

$$= \sum_{j=1}^{n}\sum_{x_j,pa_j^{\bar{S}}} \tilde{P}(x_j|pa_j^{\bar{S}})P(pa_j^{\bar{S}})\log\frac{\tilde{P}_S(x_j|pa_j^{\bar{S}})}{P_S(x_j|pa_j^{\bar{S}})} \tag{32}$$

$$= \sum_{j=1}^{n} P(pa_j^{\bar{S}})\cdot D[\tilde{P}_S(X_j|pa_j^{\bar{S}})\|P_S(X_j|pa_j^{\bar{S}})]. \tag{33}$$

Causal influence is thus observed influence plus a correction term that quantifies the divergence between the partially observed and interventional distributions. The correction term is nonnegative since it is a weighted sum of conditional Kullback–Leibler divergences.

A.2. Decomposition into conditional relative entropies. The following result generalizes Lemma 3 to the case where S contains more than one edge. It shows that the relative entropy expression defining causal strength decomposes into a sum of conditional relative entropies, each of it referring to the conditional distribution of one of the target nodes, given its parents:

LEMMA 6 (Causal influence decomposes into a sum of expectations). *The causal influence of set of arrows S can be rewritten*

$$\mathfrak{C}_S(P) = \sum_{j \in \operatorname{trg}(S)} D\Big(P(X_j|PA_j) \Big\| \sum_{pa_j^S} P(X_j|PA_j^{\bar{S}}, pa_j^S) \cdot P_{\prod}(pa_j^S)\Big),$$

where $\operatorname{trg}(S)$ *denotes the target nodes of arrows in* S.

The result is used in the proof of Theorem 5 below.

PROOF OF THEOREM 5. Using the chain rule for relative entropy [5], we get

$$D(P\|P_S) = \sum_{j=1}^{n} D[P(X_j|PA_j)\|P_S(X_j|PA_j)] \tag{34}$$

$$\begin{aligned} &= \sum_{j=1}^{n}\sum_{pa_j} P(pa_j) D[P(X_j|pa_j)\|P_S(X_j|pa_j)] \\ &= \sum_{j\in\operatorname{trg}(S)} D[P(X_j|PA_j)\|P_S(X_j|PA_j)], \end{aligned} \tag{35}$$

where we have used that $P(X_j|PA_j) = P_S(X_j|PA_j)$ for all $j \notin \operatorname{trg}(S)$. Then the statement follows from the definition of $P_S(X_j|PA_j)$. Note that a similar statement for $D(P_S\|P)$ (i.e., swapping the roles of P and P_S) would not hold because then the weighting factor $P(pa_j)$ in (35) needed to be replaced with the factor $P_S(pa_j)$, which is sensitive even to deleting edges not targeting j. □

A.3. Proof of Theorem 5. Parts (a) and (b) follow from Lemma 6 since $\mathfrak{C}_{S_i}(P)$ is the ith summand in (35), which obviously depends on $P(X_i|PA_i)$ and $P(PA_i)$ only.

To prove part (c), we will show that the restrictions of P, P_{S_1}, P_{S_2} to the variables Z, PA_Z form a so-called Pythagorean triple in the sense of [1], that is,

$$\begin{aligned} &D[P(Z, PA_Z)\|P_{S_2}(Z, PA_Z)] \\ &\quad = D[P(Z, PA_Z)\|P_{S_1}(Z, PA_Z)] + D[P_{S_1}(Z, PA_Z)\|P_{S_2}(Z, PA_Z)]. \end{aligned} \tag{36}$$

This is sufficient because the left-hand side and the first term on the right-hand side of equation (36) coincide with $\mathfrak{C}_{S_2}$ and $\mathfrak{C}_{S_1}$, respectively, due to part (b). Note, however, that

$$D[P_{S_1}(Z, PA_Z)\|P_{S_2}(Z, PA_Z)] \neq D(P_{S_1}\|P_{S_2})$$

because we have such a locality statement only for terms of the form $D(P\|P_S)$. We therefore consistently restrict attention to Z, PA_Z and find

$$
\begin{aligned}
&D\big[P(Z, PA_Z)\|P_{S_2}(Z, PA_Z)\big]\\
&\quad = \sum_{z,pa_Z} P(z, pa_Z)\log\frac{P(z|pa_Z)}{P_{S_2}(z|pa_Z^{\bar{S}_2})}\\
&\quad = \sum_{z,pa_Z} P(z, pa_Z)\log\frac{P(z|pa_Z)}{P_{S_1}(z|pa_Z^{\bar{S}_1})} + \sum_{z,pa_Z} P(z, pa_Z)\log\frac{P_{S_1}(z|pa_Z^{\bar{S}_1})}{P_{S_2}(z|pa_Z^{\bar{S}_2})}\\
&\quad = D\big[P(Z, PA_Z)\|P_{S_1}(Z, PA_Z)\big]\\
&\qquad + \sum_{z,pa_Z} P\big(z|pa_Z^{\bar{S}_1}, pa_Z^{S_1}\big)P_{\prod}\big(pa_Z^{S_1}\big)P\big(pa_Z^{\bar{S}_1}\big)\log\frac{P_{S_1}(z|pa_Z^{\bar{S}_1})}{P_{S_2}(z|pa_Z^{\bar{S}_2})},
\end{aligned}
$$

where we have used that the sources in S_1 are jointly independent and independent of the other parents of Z. By definition of P_{S_1}, the second summand reads

$$
\sum_{z,pa_Z^{\bar{S}_1}} P_{S_1}\big(z, pa_Z^{\bar{S}_1}\big)\log\frac{P_{S_1}(z|pa_Z^{\bar{S}_1})}{P_{S_2}(z|pa_Z^{\bar{S}_2})} = D\big[P_{S_1}(Z, PA_Z)\|P_{S_2}(Z, PA_Z)\big],
$$

which proves (36).

By Lemma 6, it is only necessary to prove part (d) in the case where both S and T consist of arrows targeting a single node. To keep the exposition simple, we consider the particular case of a DAG containing three nodes X, Y, Z where $S = \{X \to Z\}$ and $T = \{X \to Z, Y \to Z\}$. The more general case follows similarly. Observe that $D(P\|P_T) = 0$ if and only if

$$
P(Z|x, y) = \sum_{\hat{x},\hat{y}} P(Z|\hat{x}, \hat{y})P(\hat{x})P(\hat{y}) \tag{37}
$$

for all x, y such that $P(x, y) > 0$. Multiplying both sides with $P(x')$ and summing over all x' yields

$$
\sum_{x'} P\big(Z|x', y\big)P\big(x'\big) = \sum_{\hat{x},\hat{y}} P(Z|\hat{x}, \hat{y})P(\hat{x})P(\hat{y}),
$$

because the right-hand side does not depend on x. Using (37) again, we obtain

$$
\sum_{x'} P\big(Z|x', y\big)P\big(x'\big) = P(Z|x, y)
$$

for all x, y with $P(x, y) \neq 0$. Hence $P_S = P$, and thus $D(P\|P_S) = 0$.

A.4. Causal influence measures controllability. Causal influence is intimately related to control. Suppose an experimenter wishes to understand interactions between components of a complex system. For the causal DAG in Figure 1(d), she is able to observe nodes Y and Z, and manipulate node X. To what extent can she control node Y? The notion of control has been formalized information-theoretically in [22]:

DEFINITION 7 (Perfect control). Node Y is *perfectly controllable* by node X at $Z = z$ if, given z,

(i) states of Y are a deterministic function of states of X; and
(ii) manipulating X gives rise to all states of Y.

Perfect control can be elegantly characterized:

THEOREM 7 (Information-theoretic characterization of perfect controllability). *A node Y with inputs X and Z is perfectly controllable by X alone for $Z = z$ iff there exists a Markov transition matrix $R(x|z)$ such that*

$$\text{(C1)} \qquad H\bigl(Y|z, do(x)\bigr) := \sum_x R(x|z) H\bigl(Y|z, do(x)\bigr) = 0 \quad \textit{and}$$

$$\text{(C2)} \qquad \sum_{x \in X} P\bigl(y|z, do(x)\bigr) R(x|z) \neq 0 \qquad \textit{for all } y.$$

Here, $H(Y|z, do(x))$ denotes the conditional Shannon entropy of Y, given that $Z = z$ has been observed and X has been set to x.

PROOF. The theorem restates the criteria in the definition. For a proof, see [22]. □

It is instructive to compare Theorem 7 to our measure of causal influence. The theorem highlights two fundamental properties of perfect control. First, (C1), perfect control requires there is no variation in Y's behavior—aside from that due to the manipulation via X—given that z is observed. Second, (C2), perfect control requires that all potential outputs of Y can be induced by manipulating node X. This suggests a measure of the *degree* of control should reflect (i) the variability in Y's behavior that cannot be eliminated by imposing X values and (ii) the size of the repertoire of behaviors that can be induced on the target by manipulating a source.

For the DAG under consideration, Theorem 4 states that

$$\mathfrak{C}_{X \to Y}(P) = I(X; Y|Z) = H(Y|Z) - H(Y|X, Z).$$

The first term, $H(Y|Z)$, quantifies size of the repertoire of outputs of Y averaged over manipulations of X. It corresponds to requirement (C2) in the characterization of perfect control: that $P(y|z) > 0$ for all z. Specifically, the causal influence, interpreted as a measure of the degree of controllability, increases with the size of the (weighted) repertoire of outputs that can be induced by manipulations.

The second term, $H(Y|X, Z)$ [which coincides with $H(Y|Z, do(X))$ here], quantifies the variability in Y's behavior that cannot be eliminated by controlling X. It corresponds to requirement (C1) in the characterization of perfect control: that remaining variability should be zero. Causal influence increases as the variability $H(Y|Z, do(X)) = \sum_z P(z)H(Y|z, do(X))$ tends toward zero provided that the first term remains constant.

Acknowledgement. We are grateful to Gábor Lugosi for a helpful hint for the proof of Lemma 1 in Supplement S.1 and to Philipp Geiger for several corrections.

SUPPLEMENTARY MATERIAL

Supplement to "Quantifying causal influences" (DOI: 10.1214/13-AOS1145SUPP; .pdf). Three supplementary sections: (1) Generating an i.i.d. copy via random permutations; (2) Another option to define causal strength; and (3) The problem of defining total influence.

REFERENCES

[1] AMARI, S. and NAGAOKA, H. (1993). *Methods of Information Geometry*. Oxford Univ. Press, New York.

[2] AVIN, C., SHPITSER, I. and PEARL, J. (2005). Identifiability of path-specific effects. In *Proceedings of the International Joint Conference in Artificial Intelligence, Edinburgh, Scotland* 357–363. Professional Book Center, Denver.

[3] AY, N. and KRAKAUER, D. (2007). Geometric robustness and biological networks. *Theory in Biosciences* **125** 93–121.

[4] AY, N. and POLANI, D. (2008). Information flows in causal networks. *Adv. Complex Syst.* **11** 17–41. MR2400125

[5] COVER, T. M. and THOMAS, J. A. (1991). *Elements of Information Theory*. Wiley, New York. MR1122806

[6] GRANGER, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37** 424–38.

[7] HOLLAND, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. In *Sociological Methodology* (C. Clogg, ed.) 449–484. American Sociological Association, Washington, DC.

[8] HOYER, P., JANZING, D., MOOIJ, J., PETERS, J. and SCHÖLKOPF, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems* 21: 22*nd Annual Conference on Neural Information Processing Systems* 2008 (D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds.) 689–696. Curran Associates, Red Hook, NY.

[9] JANZING, D., BALDUZZI, D., GROSSE-WENTRUP, M. and SCHÖLKOPF, B. (2013). Supplement to "Quantifying causal influences." DOI:10.1214/13-AOS1145SUPP.

[10] LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991

[11] LEWONTIN, R. C. (1974). Annotation: The analysis of variance and the analysis of causes. *American Journal Human Genetics* **26** 400–411.

[12] MASSEY, J. (1990). Causality, feedback and directed information. In *Proc. 1990 Intl. Symp. on Info. Th. and Its Applications*. Waikiki, Hawaii.

[13] NORTHCOTT, R. (2008). Can ANOVA measure causal strength? *The Quaterly Review of Biology* **83** 47–55.

[14] PEARL, J. (2000). *Causality*: *Models*, *Reasoning*, *and Inference*. Cambridge Univ. Press, Cambridge. MR1744773

[15] PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the* 17*th Conference on Uncertainty in Artificial Intelligence* (*UAI*2001) 411–420. Morgan Kaufmann, San Francisco, CA.

[16] PÉREZ-CRUZ, F. (2009). Estimation of information theoretic measures for continuous random variables. In *Advances in Neural Information Processing Systems 21: 22nd Annual Conference on Neural Information Processing Systems 2008* (D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds.) 1257–1264. Curran Associates, Red Hook, NY.

[17] PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2011). Causal inference on discrete data using additive noise models. *IEEE Transac*. *Patt*. *Analysis and Machine Int*. **33** 2436–2450.

[18] PETERS, J., MOOIJ, J., JANZING, D. and SCHÖLKOPF, B. (2001). Identifiability of causal graphs using functional models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)* 589–598. AUAI Press, Corvallis, OR. Available at http://uai.sis.pitt.edu/papers/11/p589-peters.pdf.

[19] ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

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

[21] SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation*, *Prediction*, *and Search*. *Lecture Notes in Statistics* **81**. Springer, New York. MR1227558

[22] TOUCHETTE, H. and LLOYD, S. (2004). Information-theoretic approach to the study of control systems. *Phys*. *A* **331** 140–172. MR2046360

[23] ZHANG, K. and HYVÄRINEN, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, Montreal, Canada* 647–655. AUAI Press, Arlington, VA.

D. JANZING
M. GROSSE-WENTRUP
B. SCHÖLKOPF
MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS
SPEMANNSTR. 38
72076 TÜBINGEN
GERMANY
E-MAIL: dominik.janzing@tuebingen.mpg.de
moritz.grosse-wentrup@tuebingen.mpg.de
bernhard.schoelkopf@tuebingen.mpg.de

D. BALDUZZI
MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS
SPEMANNSTR. 38
72076 TÜBINGEN
GERMANY
AND
ETH ZÜRICH
CAB F 63.1
UNIVERSITÄTSTRASSE 6
8092 ZURICH
SWITZERLAND
E-MAIL: david.balduzzi@inf.ethz.ch