Higher-order in-and-outeractions reveal synergy and logical dependence beyond Shannon-information

Abel Jansma[1,2]

[1] *MRC Human Genetics Unit, Institute of Genetics & Cancer, University of Edinburgh*

[2] *Higgs Centre for Theoretical Physics, School of Physics & Astronomy, University of Edinburgh*

May 8, 2022

Abstract

Information theoretical quantities reveal dependencies among variables in the structure of joint, marginal, and conditional entropies, but leave some fundamentally different systems indistinguishable. Furthermore, there is no consensus on how to construct and interpret a higher-order generalisation of mutual information (MI). In this manuscript, we show that a recently proposed model-free definition of higher-order interactions (MFIs) amongst binary variables, like mutual information, is a Möbius inversion on a Boolean algebra, but of surprisal instead of entropy. This gives an information-theoretical interpretation to the MFIs, and by extension to Ising interactions. We study the dual objects to MI and MFIs on the order-reversed lattice, and find that dual MI is related to the previously studied differential mutual information, while dual interactions (outeractions) are interactions with respect to a different background state. Unlike (dual) mutual information, in- and outeractions uniquely identify all six 2-input logic gates, the dy- and triadic distributions, and different causal dynamics that are identical in terms of their Shannon-information content.

1 Introduction

> *The elementary unit of information is a difference which makes a difference*
>
> — Gregory Bateson, Steps to an Ecology of Mind [3]

In this manuscript, we investigate the role of higher-order dependencies between variables. By higher-order, we mean any structure of the system that is inherently a property of more than two variables, and that cannot be decomposed into pairwise quantities. The reason we are interested in higher-order structure is twofold. First of all, higher-order dependence corresponds to a fundamentally different kind of communication between the components of a system. If a system contains higher-order dependencies, then its structure cannot be represented by a graph, but requires a hypergraph, where a single edge can connect more than two nodes. We would like to be able to detect, and describe such systems accurately. Second, higher-order interactions play an important role in Nature, and they have been identified in genetic [22, 33, 20, 34], neuronal [26, 31, 12, 37, 13], ecological

[28, 16, 23], drug interaction[30], social [1, 6, 15], and physical [24, 7] interaction networks.

Higher-order information Interacting parts of a system are dependent in the sense that they contain or exchange shared information. It should come as no surprise, then, that information theory is one of the most successful approaches to describing dependency structures. The central questions in information theory are thus of dependence between variables. Two variables A and B with joint probability distribution $p(A, B)$ are *independent* iff for all realisations $A = a$ and $B = b$, we have that

$$p(A = a, B = b) = p(A = a)p(B = b) \tag{1}$$

Two variables are *dependent* insofar as this equation does not hold. Quantifying the dependence thus comes down to the question: *How different is the joint distribution from the product of the marginals?* The most canonical way to quantify the difference between two distributions is the KL-divergence, defined for two distributions p and q over a continuous variable X with support $\mathcal{X}$ as

$$D_{\text{KL}}\left(p(X)||q(X)\right) = \int_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}\,dx \tag{2}$$

Indeed, the KL-divergence of the joint and the product of the marginals is of central importance in information theory, and called the Mutual Information (MI):

$$\begin{aligned} MI(X, Y) &= D_{\text{KL}}\left(p(X, Y)||p(X)p(Y)\right) && (3)\\ &= \int_{y\in\mathcal{Y}}\int_{x\in\mathcal{X}} p(x, y)\log\frac{p(x, y)}{p(x)p(y)}\,dxdy && (4)\\ &= H(X) + H(Y) - H(X, Y) && (5)\end{aligned}$$

where $H(S)$ is the entropy of variables S. This immediately suggests a generalisation to dependence among a set $S = \{X_0, \ldots, X_n\}$ of n variables as:

$$MI(S) = D_{\text{KL}}\left(p(S)\ ||\ \prod_{i=0}^{n} p(X_i)\right) \tag{6}$$

For $|S| > 2$, this is known as the total correlation[32]. However, this is far from the only way to generalise dependency to higher orders. Another common choice is called the *multiple/multivariate mutual information, co-information*, or *interaction information*, and is defined on three variables as follows:

$$\begin{aligned} MI(X, Y, Z) = &H(X) + H(Y) + H(Z)\\ &-H(X, Y) - H(X, Z) - H(Y, Z))\\ &+H(X, Y, Z) && (7)\end{aligned}$$

This is symmetric in X, Y, and Z, and takes into account the pairwise entropies, but can be negative, which is not an intuitive property of information.

One particular problem of interest that the total correlation and mutual information do not address is that of synergy and redundancy. Given a set of variables with an nth-order dependency, what part of that is exclusively nth-order (called the synergistic part), and what part can also be found in a subset of $m < n$ variables (the redundant part)? Quantifying exactly to what extent shared information is synergistic is an open problem, but best addressed in the *partial information decomposition* [36], which has been applied mainly in the context of theoretical neuroscience [35]. In this manuscript, we will approach higher-order dependence in a different way, coming from a more statistical direction, but ultimately connecting our approach to these quantities from information theory.

Maximum entropy interactions In 1957, E.T. Jaynes famously showed that statistical equilibrium mechanics can be seen as a maximum entropy solution to the inverse problem of constructing a probability distribution that best reproduces a sample distribution. More precisely, the equilibrium dynamics of the (inhomogeneous, or glass-like) generalised Ising model with interactions up to nth order arise naturally as the maximum entropy distribution compatible with a data set after observing the first n moments amongst binary variables [18]. This means that to reproduce the moments in the data in a maximally non-committal way, one has to introduce higher-order interactions, *i.e.* terms that involve more than two variables, in the description of the system. Fitting such a generalised Ising model to data is nontrivial: While the log-likelihood of the Ising model is concave in the the coupling parameters, the cost of evaluating it is exponential in the total number of parameters N, so often intractable in practice [25]. In [4], the authors introduced an estimator of model-free interactions (MFIs) that exactly coincides with the solution to the inverse generalised Ising problem. Moreover, the cost of estimating all nth order interaction among N variables from M observations scales as $\mathcal{O}\left(M \cdot \binom{N}{n}\right) = \mathcal{O}(MN^n)$, i.e. polynomially[1], in the total system size N. The definition of MFIs offered in [4] seems to be a general one: besides offering a solution to the inverse generalised Ising problem, the MFIs are also expressible in terms of average treatment effects (ATEs) or regression coefficients.

We will start by restating the definition of MFIs in Section 2 and some of its properties. In Section 3, we will look more closely at the definition of MFIs and investigate its relationship to quantities from information theory. We will see that MFIs are expressible in terms of self-information, or surprisal, in the same way that mutual information can be expressed as entropies. By defining the MFIs as Möbius transformations on a lattice, we are naturally led to consider their order-theoretic dual, which we call outeractions. The order-theoretic dual of mutual information will turn out to be the differential mutual information already defined in [10]. This creates the following analogies:

- *Model-free interactions are to surprisal as mutual information is to marginal entropy.*
- *Model-free outeractions are to surprisal as differential mutual information is to marginal entropy.*

Then, in Section 4, we will show through some examples that MFIs differentiate distributions where entropy-based quantities cannot.

2 Model-free interactions

Let us start by re-defining the interactions introduced in [4] (there, these are called the *multiplicative* interactions). Define the isolated effect, or 1-point interaction, $I_i^{(Y)}$ of a

[1] This is only true in the infinite data limit. To estimate interactions on finite data, it is sometimes necessary to estimate the conditional dependencies in the data, which in the worst case scales exponentially in N again.

variable $X_i \in X$ on an observable Y as

$$I_i^{(Y)} = \frac{\partial Y}{\partial X_i}\Big|_{\underline{X}=0} \quad , \quad \underline{X} = X \setminus \{X_i\} \tag{8}$$

Where we isolate the effect of X_i on Y by conditioning on all other variables being zero. This expression is well-defined as the restriction of a derivative is the derivative of the restriction. A pair of variables X_i and X_j has a 2-point interaction $I_{ij}^{(Y)}$ when the value of X_j changes the isolated effect of X_i on Y:

$$I_{ij}^{(Y)} = \frac{\partial I_i^{(Y)}}{\partial X_j}\Big|_{\underline{X}=0} = \frac{\partial^2 Y}{\partial X_j \partial X_i}\Big|_{\underline{X}=0} \quad , \quad \underline{X} = X \setminus \{X_i, X_j\} \tag{9}$$

A third variable X_k can modulate this interaction through what we call a 3-point interaction $I_{ijk}^{(Y)}$:

$$I_{ijk}^{(Y)} = \frac{\partial I_{ij}^{(Y)}}{\partial X_k}\Big|_{\underline{X}=0} = \frac{\partial^3 Y}{\partial X_k \partial X_j \partial X_i}\Big|_{\underline{X}=0} \quad , \quad \underline{X} = X \setminus \{X_i, X_j, X_k\} \tag{10}$$

This process of taking derivatives with respect to an increasing number of variables can be repeated to define n-point interactions:

Definition 1. *(n-point interaction with respect to outcome Y) Let p be a probability distribution over a set X of random variables X_i. Let Y be a function $Y : X \to \mathbb{R}$. Then the n-point interaction $I_{X_1 \dots X_n}$ between variables $\{X_i, \dots, X_n\} \subseteq X$ is given by*

$$I_{X_1 \dots X_n}^{(Y)} = \frac{\partial^n Y(X)}{\partial X_1 \dots \partial X_n}\Big|_{\underline{X}=0} \tag{11}$$

where $\underline{X} = X \setminus \{X_1, \dots X_n\}$.

This definition of interaction makes explicit the fact that interactions are defined with respect to some outcome. We follow [4] and define interactions with respect to the most general outcome: the (log of the) joint distribution $p(X)$ over all variables X.

Definition 2. *(model-free n-point interaction between binary variables) A model-free n-point interaction (MFI) is an n-point interaction between binary random variables with respect to the logarithm of their joint probability:*

$$I_{X_1 \dots X_n} := I_{X_1 \dots X_n}^{(\log p(X))} = \frac{\partial^n \log p(X)}{\partial X_1 \dots \partial X_n}\Big|_{\underline{X}=0} \tag{12}$$

where $\underline{X} = X \setminus \{X_1, \dots X_n\}$.

In [4], the authors note that when all variables $X_i \in X$ are binary, n-point interactions become model-free in the sense that they are ratios of probabilities that do not involve the functional form of the joint probability distribution:

$$I_i = \frac{\partial \log p(X)}{\partial X_i}\Big|_{\underline{X}=0} = \log \frac{p\Big(X_i = 1 \mid \underline{X} = 0\Big)}{p\Big(X_i = 0 \mid \underline{X} = 0\Big)} \tag{13}$$

$$I_{ij} = \frac{\partial^2 \log p(X)}{\partial X_j \partial X_i}\Big|_{\underline{X}=0} = \log \frac{p\Big(X_i = 1, X_j = 1 \mid \underline{X} = 0\Big)}{p\Big(X_i = 0, X_j = 1 \mid \underline{X} = 0\Big)} \frac{p\Big(X_i = 0, X_j = 0 \mid \underline{X} = 0\Big)}{p\Big(X_i = 1, X_j = 0 \mid \underline{X} = 0\Big)} \tag{14}$$

$$\begin{aligned} I_{ijk} = \frac{\partial^3 \log p(X)}{\partial X_k \partial X_j \partial X_i}\Big|_{\underline{X}=0} = \\ \log \frac{p\Big(X_i = 1, X_j = 1 X_k = 1 \mid \underline{X} = 0\Big)}{p\Big(X_i = 0, X_j = 0, X_k = 0 \mid \underline{X} = 0\Big)} \frac{p\Big(X_i = 1, X_j = 0 X_k = 0 \mid \underline{X} = 0\Big)}{p\Big(X_i = 0, X_j = 1, X_k = 1 \mid \underline{X} = 0\Big)} \\ \times \frac{p\Big(X_i = 0, X_j = 1 X_k = 0 \mid \underline{X} = 0\Big)}{p\Big(X_i = 1, X_j = 0, X_k = 1 \mid \underline{X} = 0\Big)} \frac{p\Big(X_i = 0, X_j = 0 X_k = 1 \mid \underline{X} = 0\Big)}{p\Big(X_i = 1, X_j = 1, X_k = 0 \mid \underline{X} = 0\Big)} \end{aligned} \tag{15}$$

Here we used Bayes' rule to replace joint with conditional probabilities. This definition of interaction has the following properties:

- It is symmetric in the variables: $I_S = I_{\pi(S)}$ for any set of variables S, and any permutation π.
- Conditionally independent variables do not interact: $X_i \perp\!\!\!\perp X_j \mid \underline{X} \implies I_{ij} = 0$.
- If $\underline{X} = \emptyset$, the definition coincides with that of a log-odds ratio, which has already been considered as a measure of interaction in e.g. [14] and [2].
- The interactions are model-free: no knowledge of the functional form of $p(X)$ is required, and the probabilities can be directly estimated from i.i.d. samples.
- The MFIs are exactly the Ising interactions in the maximum entropy model after observing moments of the data. This can be readily verified by setting $p(s) = \mathcal{Z}^{-1} \exp(\sum_n \sum_{i_0,\dots,i_n} J_{1\dots n} s_1 \dots s_n)$ and using Def. 2.

Furthermore, in section A.1 we introduce and prove the following two properties of MFIs that help with the practical estimation of MFIs:

- An n-point interaction can only be nonzero if all n variables are in each other's minimal Markov blanket.
- If $\underline{X}$ does not include the full complement of the interacting variables, the bias this induces in the estimate of the interaction is proportional to the pointwise mutual information of states where the omitted variables are 0.

3 Information theory and Möbius inversions

3.1 Mutual information as a Möbius inversion

The definition of an n-point interaction as a derivative of a derivative is reminiscent of Gregory Bateson's view of information as a *difference which makes a difference* [3], but the relationship between information theory and the model-free interactions is more than linguistic. It turns out that interaction and information are generalised derivatives of similar functions on Boolean algebras. To see this, consider the definition of mutual information, and its higher-order generalisations:

$$\begin{tikzcd}
& \{X,Y\}=\hat{1} \arrow[dl] \arrow[dr] & \\
\{X\} \arrow[dr] & & \{Y\} \arrow[dl] \\
& \emptyset=\hat{0} &
\end{tikzcd}$$

$$\begin{tikzcd}
& \{X,Y,Z\}=\hat{1} \arrow[dl] \arrow[d] \arrow[dr] & \\
\{X,Y\} \arrow[d] \arrow[dr] & \{X,Z\} \arrow[dl] \arrow[dr] & \{Y,Z\} \arrow[dl] \arrow[d] \\
\{X\} \arrow[dr] & \{Y\} \arrow[d] & \{Z\} \arrow[dl] \\
& \emptyset=\hat{0} &
\end{tikzcd}$$

Figure 1: The lattice associated to $\mathcal{P}(\{X,Y\})$ (left) and $\mathcal{P}(\{X,Y,Z\})$ (right), ordered by inclusion. An arrow $b \to a$ indicates $a < b$.

- Mutual information:

$$\begin{aligned} MI(X,Y) &= H(X) - H(X \mid Y) && (16)\\ &= H(X) + H(Y) - H(X,Y) && (17) \end{aligned}$$

- The generalisation of mutual information that describes higher-order dependencies between variables goes by many names: *multiple mutual information, co-information*, or *interaction information*, and is defined on three variables as follows:

$$\begin{aligned} MI(X,Y,Z) =& MI(X,Y) - MI(X,Y \mid Z) && (18)\\ =& H(X) + H(Y) + H(Z) \\ -& H(X,Y) - H(X,Z) - H(Y,Z)) \\ +& H(X,Y,Z) && (19) \end{aligned}$$

I will refer to both these quantities simply as mutual information, defined by a so-called Möbius inversion.

Möbius inversions Each MI-based quantity can be written as a specific sum of marginal entropies of subsets of the set of variables. Given a finite set of variables S, its powerset $\mathcal{P}(S)$ can be given a partial ordering as follows:

$$a \leq b \iff a \subseteq b \quad \forall\, a,b \in \mathcal{P}(S) \tag{20}$$

This poset $P = (\mathcal{P}(S), \subseteq)$ is called a Boolean algebra, and since each pair of sets has a unique supremum (their union) and infimum (their intersection), it is a lattice. This lattice structure is visualised for two and three variables in figure 1. In general, the lattice of an n-variable Boolean algebra forms an n-cube.

On a poset P, define the Möbius function $\mu_P : P \times P \to \mathbb{R}$ as[2]

[2] This function type makes μ_P an element of the *incidence algebra* of P. In fact, μ is the inverse of the zeta function $\zeta : \zeta(x,y) = 1$ iff $x \leq y$, 0 otherwise.

$$\mu_P(x,y) = \begin{cases} 1 & \text{if } x = y \\ -\sum\limits_{z: x \leq z < y} \mu_P(x,z) & \text{if } x < y \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

On a powerset ordered by inclusion, the Möbius function takes the simple form $\mu(x,y) = (-1)^{|x|-|y|}$ [29, 27]. This definition allows us to write the mutual information among a set of variables τ as [5, 11]:

$$MI(\tau) = (-1)^{|\tau|-1} \sum_{\eta \leq \tau} \mu_P(\eta, \tau) H(\eta) \tag{22}$$

$$= \sum_{\eta \leq \tau} (-1)^{|\eta|+1} H(\eta) \tag{23}$$

Where P is the Boolean algebra with $\tau = \hat{1}$, and $H(\eta)$ is the marginal entropy of the set of variables η. This indeed coincides with equation 17 for $\tau = \{X, Y\}$ and equation 19 for $\tau = \{X, Y, Z\}$. Equation 22 is a convolution known as a Möbius inversion:

Definition 3. *(Möbius inversion over a poset, Rota 1964 [27]) Let P be a poset* $(S, \leq)$*. Let* $\mu : P \times P \to \mathbb{R}$ *be the Möbius function from equation 21. Let* $g : P \to \mathbb{R}$ *be a function on* P*. Then the function*

$$f(y) = \sum_{x \leq y} \mu_P(x,y) g(x) \tag{24}$$

is called the Möbius inversion of g *on* P*. Furthermore, this equation can be inverted to yield*

$$f(y) = \sum_{x \leq y} \mu_P(x,y) g(x) \iff g(y) = \sum_{x \leq y} f(x) \tag{25}$$

The Möbius inversion is a generalisation of the derivative to posets. If $P = (\mathbb{N}, \leq)$, equation 25 is just a discrete version of the fundamental theorem of calculus [29]. Equation 25 also implies that we can express joint entropy as a sum over mutual information:

$$H(\tau) = (-1)^{|\tau|-1} \sum_{\eta \leq \tau} MI(\eta) \tag{26}$$

For example, in the case of three variables:

$$H(X,Y,Z) = MI(X,Y,Z) + MI(X,Y) + MI(X,Z) + MI(Y,Z) + H(X) + H(Y) + H(Z) \tag{27}$$

Instead of starting with entropy, we can also start with a quantity known as surprisal, or self-information, defined as the log of the probability of a certain state:

$$S(X = x) = -\log p(X = x) \tag{28}$$

Which, with standard abuse of notation, can simply be written as:

$$S(x) = -\log p(x) \tag{29}$$

Surprisal plays an important role in information theory, and indeed, the expected surprisal across all possible realisations $X = x$ is the entropy of the variables X:

$$\mathbb{E}_X[S(X = x)] = H(X) \tag{30}$$

As we are often interested in the marginal surprisal of a realisation $X = x$, summed over Y, let us write this explicitly as

$$\log p(x; Y) \coloneqq \sum_y \log p(x, y) \tag{31}$$

With this, consider the Möbius inversion of the marginal surprisal over the lattice P:

$$\text{pmi}(T = \tau) \coloneqq (-1)^{|\tau|} \sum_{\eta \leq \tau} \mu_P(\eta, \tau) \log p(\eta; \tau \setminus \eta) \tag{32}$$

This is a generalised version of the pointwise mutual information, usually defined on just two variables:

$$\begin{align}
\text{pmi}(X = x, Y = y) &= \log(x, y; \emptyset) - \log(x; Y) - \log(y; X) + \log(\emptyset; X, Y) \tag{33}\\
&= \log \frac{p(x, y)}{p(x)p(y)} \tag{34}
\end{align}$$

Summary

- *Mutual information is the Möbius inversion of marginal entropy.*
- *Pointwise mutual information is the Möbius inversion of marginal surprisal.*

3.2 MFIs as a Möbius inversion

Now that we have defined mutual information in terms of Möbius inversions, we can do the same for the model-free interactions. We start, again, with (negative) surprisal. However, on Boolean variables, a state is just a partition of the variables into two sets: one where the variables are set to 1, and one where they are set to 0. That means we can write the surprisal of observing a particular state by just specifying which variables $X \subseteq Z$ are set to 1, keeping all other variables $Z \setminus X$ at 0, which we will write as:

$$S_{X;Z} \coloneqq \log p(X = 1, Z \setminus X = 0) \tag{35}$$

Definition 4. *(Interactions as Möbius inversions) Let p be a probability distribution over a set T of random variables. Let $P = (\mathcal{P}(\tau), \subseteq)$, the powerset of a set $\tau \subseteq T$ ordered by inclusion. Then the interaction $I(\tau; T)$ among variables τ is given by*

$$\begin{align}
I(\tau; T) \coloneqq& \sum_{\eta \leq \tau} \mu_P(\eta, \tau) S_{\eta;T} \tag{36}\\
=& \sum_{\eta \leq \tau} (-1)^{|\eta| - |\tau|} \log p(\eta = 1, T \setminus \eta = 0) \tag{37}
\end{align}$$

For example, when τ contains a single variable $X \subseteq T$, then

$$I(\{X\};T) = \mu_P(\{X\},\{X\})S_{\{X\};T} + \mu_P(\emptyset,\{X\})S_{\emptyset;T} \tag{38}$$

$$= \log\frac{p(X=1, T\setminus X=0)}{p(X=0, T\setminus X=0)} \tag{39}$$

Which coincides with the 1-point interaction in equation 13. Similarly, when τ contains two variables $\tau = \{X,Y\} \subseteq T$, then

$$I(\{X,Y\};T) = \mu_P(\{X,Y\},\{X,Y\})S_{\{X,Y\};T} + \mu_P(\{X\},\{X,Y\})S_{\{X\};T} \tag{40}$$

$$+\mu_P(\{Y\},\{X,Y\})S_{\{Y\};T} + \mu_P(\emptyset,\{X,Y\})S_{\emptyset;T}$$

$$= \log\frac{p(X=1,Y=1,T\setminus\{X,Y\}=0)p(X=0,Y=0,T\setminus\{X,Y\}=0)}{p(X=1,Y=0,T\setminus\{X,Y\}=0)p(X=0,Y=1,T\setminus\{X,Y\}=0)} \tag{41}$$

Which coincides with the 2-point interaction in equation 14. In fact, this pattern holds in general:

Theorem 1. *(Equivalence of interactions) The interaction $I(\tau,T)$ from Definition 4 is the same as the model-free interaction I_τ from Definition 2. That is, for any set of variables $\tau \subseteq T$*

$$I(\tau,T) = I_\tau \tag{42}$$

Proof. We have to show that

$$\sum_{\eta\leq\tau}(-1)^{|\eta|-|\tau|}\log p(\eta=1,T\setminus\eta=0) = \frac{\partial^n \log p(X)}{\partial X_1\ldots\partial X_n}\Big|_{\underline{X}=0} \tag{43}$$

Both sides of this equation are sums of $\pm\log p(s)$, where s is some binary string, so we have to show that the same strings appear with the same sign.

First, note that the Boolean algebra of sets ordered by inclusion (as in figure 1), is equivalent to the poset of binary strings where for any two strings a and b, $a \leq b \iff a\wedge b = a$. The equivalence follows immediately upon setting each element $a \in \mathcal{P}(S)$ to the string where $a=1$ and $S\setminus a = 0$. This map is one-to-one and monotonic with respect to the partial order as $A\subseteq B \iff A\cap B = A$. That means we can rewrite Definition 4 as a Möbius inversion on the lattice of Boolean strings $S = (\mathbb{B}^{|\tau|},\leq)$ (shown for the 3-variable case on the left side of figure 2):

$$I(\tau;T) = \sum_{s\leq\hat{1}_S}\mu_S(s,\hat{1}_S)\log p(\tau=s,T\setminus\tau=0) \tag{44}$$

Note that for any pair (α,τ) where $\alpha\subseteq\tau$, with respective string representations $(s,t)\in\mathbb{B}^{|\tau|}\times\mathbb{B}^{|\tau|}$, we have the following:

$$|\tau|-|\alpha| = \sum_i (t\wedge\neg s)_i \tag{45}$$

So that we can write

$$I(\tau;T) = \sum_{s\leq\hat{1}_S}(-1)^{\sum\neg s}\log p(\tau=s,T\setminus\tau=0) \tag{46}$$

To see that this is exactly the boolean derivative from definition 2, define a map

$$e^{(n)}_{i,s} : \mathcal{F}_{\mathbb{B}^n} \to \mathcal{F}_{\mathbb{B}^{n-1}} \tag{47}$$

where $\mathcal{F}_{\mathbb{B}^n}$ is the set of functions from n Boolean variables to $\mathbb{R}$. This map is defined as

$$e^{(n)}_{i,s} : f(X_0, \dots X_i, \dots X_n) \mapsto f(X_0, \dots X_i = s, \dots X_n) \tag{48}$$

With this map, the Boolean derivative of a function $f(X_0, \dots, X_n)$ can be written as

$$\frac{\partial}{\partial X_i} f(X) = (e^{(n)}_{i,1} - e^{(n)}_{i,0}) f(X) \tag{49}$$

$$= f(X_1, \dots, X_i = 1, \dots, X_n) - f(X_1, \dots, X_i = 0, \dots, X_n) \tag{50}$$

Such that the derivative w.r.t. a set S of m variables becomes function composition:

$$\left(\prod_{i=0}^{m} \frac{\partial}{\partial X_{S_i}} \right) f(X) = \left(\bigcirc_{i=0}^{m} (e^{(n-i)}_{S_i,1} - e^{(n-i)}_{S_i,0}) \right) f(X) \tag{51}$$

From this, it is clear that a term $f(s)$ appears with a minus sign iff $e^{(n)}_{i,0}$ has been applied an odd number of times. Therefore, terms where s contains an odd number of 0s get a minus sign. This can be summarised as:

$$\left(\prod_{i=0}^{m} \frac{\partial}{\partial X_{S_i}} \right) f(X) = \sum_{s \in \mathbb{B}^n} (-1)^{\sum \neg s} f(X_S = s, X \setminus X_S) \tag{52}$$

We can therefore write

$$I_\tau = \sum_{s \in \mathbb{B}^n} (-1)^{\sum \neg s} \log(\tau = s, T \setminus \tau = 0) \tag{53}$$

Noting that the sums $\sum_{s \leq \hat{1}_S}$ and $\sum_{s \in \mathbb{B}^n}$ contain exactly the same terms equates equations 53 and 46, and completes the proof.

□

Note that the structure of the lattice S reveals some structure in the interactions that was already noted in [4]. On the right side of figure 2, we show the 3-variable lattice again, this time shading two regions. The green region corresponds to the 2-point interaction between the first two variables. The red region is a similar interaction between the first two variables, but this time in the context of the third variable fixed to 1, instead of 0. This shows the interpretation of a 3-point interaction as the difference in two 2-point interactions: $I_{XYZ} = I_{XY}|_{Z=1} - I_{XY}$. The symmetry of the hypercube shows the three different but equivalent choices for which variable to set to 1. Treating the Boolean algebra as a die where the sides facing up are ⚀, ⚁, and ⚂, we have that

$$I_{XYZ} = \text{⚀} - \text{⚅} = \text{⚁} - \text{⚄} = \text{⚂} - \text{⚃} \tag{54}$$

As before, we can invert Definition 4 and express the surprise of observing a state with 1s in terms of interactions:

$$\log p(\tau = 1, T \setminus \tau = 0) = \sum_{\eta \leq \tau} I(\eta, T) \tag{55}$$

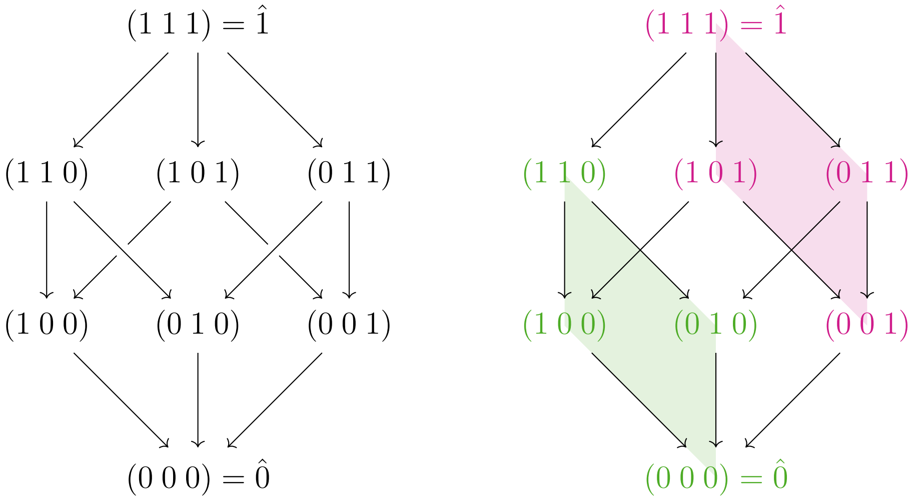


Figure 2:
Left: The lattice associated to $\mathcal{P}(\{X, Y, Z\})$, ordered by inclusion, as binary strings. Equivalently, the lattice of binary strings, where for any two strings a and b, $a \leq b \iff a \wedge b = a$.
Right: Two regions are shaded, corresponding to the decomposition of the 3-point interaction into two 2-point interactions.

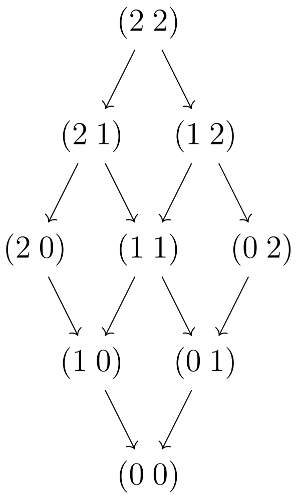


Figure 3: The lattice of two variables that can take three values, ordered by $a \leq b \iff \forall i : a_i \leq b_i$.

For example, in the case where $T = \{X, Y, Z\}$ and $\tau = \{X, Y\}$

$$S(1,1,0) = -\log p(1,1,0) = -I_{XY} - I_X - I_Y - I_\emptyset \tag{56}$$

Which illustrates that when X and Y tend to be off ($I_X < 0$ and $I_Y < 0$), and X and Y tend to be different ($I_{XY} < 0$), then observing the state (1, 1, 0) is very suprising.

Categorical interactions If we take the definition of interactions as the Möbius inversion of surprisal seriously, we might ask the question what happens when instead of using a Boolean algebra, we invert surprisal over a different lattice. One example is shown in figure 3, and it corresponds to variables that can take 3 values – 0, 1, or 2 – where states are ordered by $a \leq b \iff \forall i : a_i \leq b_i$. To calculate interactions on this lattice, we need to know the value of Möbius functions of the type $\mu(s, 22)$. It can be readily verified that most Möbius functions like this are zero, except for $\mu(22, 22) = \mu(11, 22) = 1$, and $\mu(21, 22) = \mu(12, 22) = -1$, which gives us exactly the terms in the interactions between two categorical variables changing from $1 \to 2$, as defined in [4]. Calculating interactions of different sublattices with $\hat{1} = (21), (12)$, or (11) gives us the other categorical interactions. The transitivity property of the interactions, i.e. $I(X : 0 \to 2, Y : 0 \to 1) = I(X : 0 \to 1, Y : 0 \to 1) + I(X : 1 \to 2, Y : 0 \to 1)$, follows immediately from the structure of the lattice in figure 3, and the alternating signs of the Möbius functions on a Boolean algebra.

3.3 Information and interactions on dual lattices

Lattices have the property that the set with the reversed order is still a lattice. That is, if $\mathcal{L} = (S, \leq)$ is a lattice, then $\mathcal{L}^{\text{op}} = (S, \preceq)$, where $\forall a, b \in S : a \preceq b \iff a \geq b$, is also a lattice. This raises the question: what corresponds to mutual information and interaction on these dual lattices[3]?

Dual information We can calculate the dual notion of mutual information, denoted MI^*, by first noting that the dual to a Boolean algebra is a Boolean algebra, so we still have that $\mu(x, y) = (-1)^{|x|-|y|}$, and simply replacing P by P^{op} in equation 22:

$$MI^*(\tau) = \sum_{\eta \preceq \tau} (-1)^{|\eta|+1} H(\eta) \tag{57}$$

The dual mutual information of $\tau = \hat{1}_{P^{\text{op}}}$ is just $MI^*(\emptyset) = MI(\hat{1}_P)$, the mutual information among all variables. However, the dual mutual information of a singleton set X is:

$$\begin{aligned} MI^*(X) &= MI(\hat{1}_P) - MI(\hat{1}_P \setminus X) && (58) \\ &= \Delta(X; \hat{1}_P \setminus X) && (59) \end{aligned}$$

where Δ is known as the differential mutual information [10]. It describes the change in mutual information when leaving out X, and has been used to describe information structures in genetics [11]. On the Boolean algebra of three variables $\{X, Y, Z\}$, the dual mutual information of X can be written out as:

[3] Recognising that a poset $\mathcal{L} = (S, \leq_{\mathcal{L}})$ is a category $\mathcal{C}$ with objects S and a morphism $f : A \to B$ iff $B \leq_{\mathcal{L}} A$, these become questions in the *opposite* category $\mathcal{C}^{\text{op}}$.

$$\begin{aligned} MI^*(X) &= \mu(\{X\},\{X\})H(X) + \mu(\{X,Y\},\{X\})H(X,Y) + \\ &\mu(\{X,Z\},\{X\})H(X,Z) + \mu(\{X,Y,Z\},\{X\})H(X,Y,Z) \qquad (60) \\ &= H(X) - H(X,Y) - H(X,Z) + H(X,Y,Z) \qquad (61) \end{aligned}$$

Since Δ is the dual of mutual information, it should arguably be called the mutual co-information, but the term co-information is unfortunately already used to refer to normal higher-order mutual information.

Outeractions To find the dual to the interactions, let's start from equation 44 and construct $S^{\text{op}} = (\mathbb{B}^{|\tau|}, \preceq)$, dual to the lattice of binary strings $S = (\mathbb{B}^{|\tau|}, \leq)$.

A dual interaction of variables $\tau \subseteq T$ will be denoted $I^*(\tau; T)$, and is defined as follows:

$$I^*(\tau;T) \coloneqq \sum_{s \preceq \hat{1}_{S^{\text{op}}}} \mu_{S^{\text{op}}}(s, \hat{1}_{S^{\text{op}}}) \log p(\tau = s, T \setminus \tau = 0) \qquad (62)$$

Again, when $\tau = \hat{1}_{S\text{op}} = \hat{0}_S = \emptyset$, this is just $(-1)^{|\tau|} I(\hat{1}_S)$, but the dual interaction of a singleton set X is:

$$I^*(X;T) = (-1)^{|\hat{1}_S|-1}\Big(I(\hat{1}_S;T) + I(\hat{1}_S \setminus X;T)\Big) \qquad (63)$$

For example, on the three variable lattice in figure 2, the dual interaction of X is

$$I^*(X;T) = I(X,Y,Z;T) + I(Y,Z;T) \qquad (64)$$

Writing p_{ijk} for $p(X = i, Y = j, Z = k \mid T \setminus \{X,Y,Z\} = 0)$, we see that this is equal to:

$$I^*(X;T) = \log \frac{p_{111}p_{100}}{p_{101}p_{110}} \qquad (65)$$

Which is similar to the 2-point interaction I_{YZ} defined in equation 14, except conditioned on $X = 1$ instead of 0. Dual interactions should probably be called co-interactions, but to avoid confusion with the term co-information (already in use to refer to higher-order mutual information) I will instead refer to the dual interactions as outeractions. Outeractions are just interactions, conditioned on certain variables being 1 instead of 0. This makes them no longer equal to the Ising interactions between Boolean variables, but there are situations in which an interaction is more interesting in the context with $Z = 1$ instead of $Z = 0$, for example if Z is always 1 in all situations of interest.

Summary

- *Mutual information is the Möbius inversion of entropy on the lattice of subsets ordered by inclusion.*
- *Differential (or conditional) mutual information is the Möbius inversion of entropy on the dual lattice.*
- *Model-free interactions are the Möbius inversion of surprisal on the lattice of subsets ordered by inclusion.*
- *Model-free outeractions are the Möbius inversion of surprisal on the dual lattice.*

- *Outeractions of a variable X are interactions between the other variables where X is set to 1 instead of 0.*

To summarise these relationships diagrammatically, note that surprisals form a vector space as follows. Let $\mathcal{P}(T)$ be the powerset of a set of variables T, and let $|\mathcal{P}(T)| = n$. This forms a lattice $P = (\mathcal{P}(T), \subseteq)$ ordered by inclusion, so $\mathcal{P}(T)$ can be given a topological ordering, indexed by i as $\mathcal{P}(T) = \cup_{i=0}^{n} t_i$. Let $\mathcal{S}$ be the set of linear combinations of surprisals of subsets of T:

$$\mathcal{S} = \left\{ \sum_{i=0}^{n} a_i \log p(t_i) \mid a_i \in \mathbb{R} \right\} \tag{66}$$

This set is given a vector space structure over $\mathbb{R}$ by the usual scalar multiplication and addition. Note that the set

$$\mathcal{B} = \{\log p(t) \mid t \in \mathcal{P}(T)\} \tag{67}$$

forms a basis for this vector space, since $\sum_i \alpha_i \log p(t_i) = 0$ has no non-trivial solutions[4], and $\text{span}(\mathcal{B}) = \mathcal{S}$. To define a map from $\mathcal{S} \to \mathbb{R}$, we only need to specify its action on $\mathcal{B}$, and extend the definition linearly. That means we can fully define the map $eval_T : \mathcal{S} \to \mathbb{R}$ by specifying:

$$eval_T : \log p(R = r) \mapsto \log p(R = 1, T \setminus R = 0) \tag{68}$$

Similarly, define the expectation map $\mathbb{E} : \mathcal{S} \to \mathbb{R}$ as

$$\mathbb{E} : \log p(R = r) \mapsto \sum_r p(R = r) \log(R = r) \tag{69}$$

which outputs the expected surprise over all realisations $R = r$. Finally, note that the Möbius inversion over a poset P is an endomorphism of the set $\mathcal{F}_P$ of functions over P, defined as

$$M_P : \mathcal{F}_P \to \mathcal{F}_P \tag{70}$$

$$M_P : f(y) \mapsto \sum_{x \leq y} \mu(x, y) f(x) \tag{71}$$

Together, these three maps make the following diagram commute:

$$\begin{tikzcd}
MI^*(R) = \Delta(R; \hat{1}_P) & H(R) \arrow[l, "M_{P^{op}}"'] \arrow[r, "M_P"] & MI(R) \\
\text{pmi}^*(R = r) \arrow[u, "\mathbb{E}"] \arrow[d, "eval_T"] & S(R = r; T) \arrow[l, "M_{P^{op}}"'] \arrow[r, "M_P"] \arrow[u, "\mathbb{E}"] \arrow[d, "eval_T"] & \text{pmi}(R = r) \arrow[u, "\mathbb{E}"] \arrow[d, "eval_T"] \\
I^*(R; T) & S_{R;T} \arrow[l, "M_{P^{op}}"'] \arrow[r, "M_P"] & I(R; T)
\end{tikzcd}$$

[4] Only when two variables a and b are independent do we have linear dependencies in $\mathcal{B}$, as then $\log p(a, b) = \log p(a) + \log p(b)$.

For the case where $T = \{X, Y, Z\}$ and $R = \{X, Y\}$, this explicitly amounts to:

$$\begin{tikzcd}
\begin{array}{c}\sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} p(x,y,z)\log p(x,y,z)\\ -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y)\end{array} & \sum_{(x,y)\in X\times Y} p(x,y)\log p(x,y) \arrow[l, "M_{Pop}"'] \arrow[r, "M_P"] & \begin{array}{c}\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y)\log p(x,y)\\ -\sum_{x\in\mathcal{X}} p(x)\log p(x)\\ -\sum_{y\in\mathcal{Y}} p(y)\log p(y)\end{array} \\
\log\frac{p(x,y,z)}{p(x,y)} \arrow[u, "\mathbb{E}"'] \arrow[d, "eval_T"'] & \log p(x,y) \arrow[l, "M_{Pop}"'] \arrow[r, "M_P"] \arrow[u, "\mathbb{E}"'] \arrow[d, "eval_T"'] & \log\frac{p(x,y)p(\emptyset)}{p(x)p(y)} \arrow[u, "\mathbb{E}"'] \arrow[d, "eval_T"'] \\
\log\frac{p(1,1,1)}{p(1,1,0)} & \log p(1,1,0) \arrow[l, "M_{Pop}"'] \arrow[r, "M_P"] & \log\frac{p(1,1,0)p(0,0,0)}{p(1,0,0)p(0,1,0)}
\end{tikzcd}$$

4 Advantages of in- and outeractions over entropy-based quantities

While mutual information and interactions are related, there are some important differences in how they capture dependencies. Note, for example, that higher-order information quantities are not independent of lower-order quantities. The mutual information of three variables is bounded by the pairwise quantities as

$$-\min\{MI(X,Y \mid Z), MI(Y,Z \mid X), MI(X,Z \mid Y)\} \leq MI(X,Y,Z) \leq \min\{MI(X,Y), MI(Y,Z), MI(X,Z)\} \tag{72}$$

That means that there are no systems with zero pairwise mutual information, but positive higher-order information. This is not true for the interactions: each interaction, at any order, is free to vary, and each combination of interactions defines a unique probability distribution, most easily identified in the way they describe the Boltzmann distribution of a generalised Ising model. Furthermore, a class of neural networks known as Restricted Boltzmann Machines are known universal approximators [9], and exactly, though not uniquely, encode the Boltzmann distribution of a generalised Ising model in one of its layers [8, 25]. That is, each distribution is uniquely determined by its set of interactions, and should be distinguishable by them. This is famously not true for entropy-based Shannon information quantities, as we will illustrate here with some examples.

4.1 In- and outeractions quantify and distinguish synergy in logic gates

Under the assumption of a causal collider structure $A \rightarrow C \leftarrow B$, nonzero 3-point interactions I_{ABC} can be interpreted as logic gates. A positive 3-point interaction means that the numerator in equation 15 is larger than the denominator. Under the sufficient but not necessary assumption that each term in the numerator is larger than each term in the denominator, we get the following truth table as $I_{ABC} \rightarrow +\infty$:

A	B	C
0	0	1
0	1	0
1	0	0
1	1	1

Which describes an XNOR gate. Let $p_{\mathcal{G}}$ be the probability of each of the four states in the truth table for a gate $\mathcal{G}$, and let $\epsilon_{\mathcal{G}}$ be the probability of all other states. Then the 3-point interaction of an XNOR gate can be written as:

$$I_{ABC}^{\text{XNOR}} = \log \frac{p_{\text{XNOR}}^4}{\epsilon_{\text{XNOR}}^4} \tag{73}$$

Similarly, from the truth tables of AND and OR gates, we get

$$I_{ABC}^{\text{AND}} = \log \frac{\epsilon_{\text{AND}}\, p_{\text{AND}}^3}{\epsilon_{\text{AND}}^3 p_{\text{AND}}} \tag{74}$$

$$I_{ABC}^{\text{OR}} = \log \frac{\epsilon_{\text{OR}}^3 p_{\text{OR}}}{\epsilon_{\text{OR}}\, p_{\text{OR}}^3} \tag{75}$$

If we consider equally noisy gates so that $p_{\mathcal{G}} = p$ and $\epsilon_{\mathcal{G}} = \epsilon$, we can directly compare the gates. Note that when a gate has a 3-point interaction I, its logical negation will have 3-point interaction $-I$. This determines the 3-point interactions of all $2^3 = 6$ possible 2-input logical gates, summarised in table 1. The two gates with the strongest absolute interactions, XNOR and XOR, are also the only two gates that are purely synergistic, i.e. knowing just one of the two inputs gives you no information about the output. This relationship to synergy holds for 3-input gates as well. The 3-input gate with the strongest 4-point interaction has the following truth table:

A	B	C	D
0	0	0	0
0	0	1	1
0	1	0	1
1	0	0	1
0	1	1	0
1	0	1	0
1	1	0	0
1	1	1	1

It is a 3-input XOR gate, i.e. $D = (A + B + C) \mod 2$, and is again maximally synergistic, since observing only 2 of the 3 inputs gives zero bits of information on the output. Setting this maximum 4-point interaction to I, the 3-input OR and AND gates get 4-point interaction $I/4$, and the 3-input XOR gate that outputs 0 on (0, 0, 0) has 4-point interaction $3I/4$, so the hierarchies of interaction and synergy still match.

The 3-point interactions are able to separate most 2-input logic gates by sign or value, leaving only AND $\sim$ NOR and OR $\sim$ NAND. Mutual information has less resolving power. Assuming a uniform distribution over all 4 allowed states from a gate's truth table, a brief calculation yields:

$$MI^{\text{OR}}(A,B,C) = MI^{\text{AND}}(A,B,C) = MI^{\text{NOR}}(A,B,C) = MI^{\text{NAND}}(A,B,C) \tag{76}$$
$$= -\log\left(\frac{3^{3/4}}{4}\right) - 1 \approx -0.189$$

$$MI^{\text{XOR}}(A,B,C) = MI^{\text{XNOR}}(A,B,C) = -1 \tag{77}$$

$\mathcal{G}$	$I^{\mathcal{G}}_{ABC}$
XNOR	I
XOR	$-I$
AND	$\frac{1}{2}I$
OR	$-\frac{1}{2}I$
NAND	$-\frac{1}{2}I$
NOR	$\frac{1}{2}I$

Table 1: The 3-point interactions for all 2-input logic gates at equal noise level are related through $I = 4\log\frac{p}{\epsilon}$, and degenerate in AND $\sim$ NOR and OR $\sim$ NAND.

$\mathcal{G}$	$H(A)$ $=H(B)$	$H(C)$	$H(A,B)$	$H(A,C)$ $=H(B,C)$	$H(A,B,C)$
XNOR	1	1	2	2	2
XOR	1	1	2	2	2
AND	1	$\log\frac{3^{3/4}}{4}$	2	$\frac{3}{2}$	2
OR	1	$\log\frac{3^{3/4}}{4}$	2	$\frac{3}{2}$	2
NAND	1	$\log\frac{3^{3/4}}{4}$	2	$\frac{3}{2}$	2
NOR	1	$\log\frac{3^{3/4}}{4}$	2	$\frac{3}{2}$	2

Table 2: The marginal entropies of variables in a logic gate are degenerate in XOR $\sim$ XNOR and AND $\sim$ OR $\sim$ NAND $\sim$ NOR.

That is, mutual information resolves strictly fewer logical gates by value, and none by sign. In fact, all entropy-based quantities inherit the degeneracy summarised in table 2.

The outeractions $I^{*\mathcal{G}}_{C} = I^{\mathcal{G}}_{ABC} + I^{\mathcal{G}}_{AB}$ contain the same degeneracy as the 3-point interactions. However, let us use the same sign-convention as differential mutual information and define a new quantity $J^{*\mathcal{G}}_{A} = I^{\mathcal{G}}_{ABC} - I^{\mathcal{G}}_{BC}$. This quantity assigns a different value to each logic gate $\mathcal{G}$. The symmetric quantity $\overline{J}^{*\mathcal{G}} = J^{*\mathcal{G}}_{A} J^{*\mathcal{G}}_{B} J^{*\mathcal{G}}_{C}$, the interaction analogous to the symmetric deltas from [10], inherits the perfect resolution from $J^{*\mathcal{G}}_{A}$. This is summarised in Table 3. The J-interactions thus uniquely assign a value to each gate, proportional to the synergy of its logic. The hierarchy is $J^{*\text{XNOR}}_{A} > J^{*\text{NOR}}_{A} > J^{*\text{AND}}_{A}$, mirrored for their logical complement. XNOR is indeed the most synergistic, and NOR is more synergistic than AND with respect to observing a 0 in one of the inputs: In a NOR gate, a 0 in the input gives no information on the output, while it completely fixes the output of an AND gate. Since the interactions are defined in a context of 0s, they order synergy with respect to observing 0s.

$\mathcal{G}$	MI_{ABC}	$I^{\mathcal{G}}_{ABC}$	$I^{*\mathcal{G}}_{A}$	$J^{*\mathcal{G}}_{A}$	$J^{*\mathcal{G}}_{C}$	$\overline{J}^{*\mathcal{G}}$
XNOR	-1	I	$\frac{1}{2}I$	$\frac{3}{2}I$	$\frac{3}{2}I$	$\frac{27}{8}I^3$
XOR	-1	$-I$	$-\frac{1}{2}I$	$-\frac{3}{2}I$	$-\frac{3}{2}I$	$-\frac{27}{8}I^3$
AND	-0.189	$\frac{1}{2}I$	$\frac{1}{2}I$	$\frac{1}{2}I$	$\frac{3}{4}I$	$\frac{3}{16}I^3$
OR	-0.189	$-\frac{1}{2}I$	0	$-I$	$-\frac{3}{4}I$	$-\frac{3}{4}I^3$
NAND	-0.189	$-\frac{1}{2}I$	$-\frac{1}{2}I$	$-\frac{1}{2}I$	$-\frac{3}{4}I$	$-\frac{3}{16}I^3$
NOR	-0.189	$\frac{1}{2}I$	0	I	$\frac{3}{4}I$	$\frac{3}{4}I^3$

Table 3: While the interactions leave some gates indistinguishable, the outeractions (with the appropriate sign convention) of the inputs are unique to each gate. As before: $I = 4\log\frac{p}{\epsilon}$.

4.2 Interactions distinguish dynamics and causal structures

To see how different association metrics reflect the underlying causal dynamics, we simulate data generated from a selection of 3-node causal DAGs. On a given DAG $\mathcal{G}$, denote the set of nodes without parents, the orphan nodes, by S_0. Each orphan node in S_0 gets a random value drawn from a Bernoulli distribution, i.e. $P(X = 1) = p$ and $P(X = 0) = 1-p$. Denote the set of children of orphan nodes as S_1. Each node in S_1 then gets set to either the product of its parent nodes (for *multiplicative* dynamics), or the mean of its parent nodes (for *additive* dynamics), plus some zero-mean Gaussian noise with variance σ^2. All nodes are then rounded to a 0 or 1. A set S_2 is then defined as the set of all children of nodes in S_1, and these get a value using the same dynamics as before. As long as the causal structure is acyclic, this algorithm terminates on a set of nodes S_i that has no children. For example, the chain graph $A \to B \to C$ has $S_0 = \{A\}$, $S_1 = \{B\}$, $S_2 = \{C\}$, and $S_3 = \emptyset$, at which point the updating terminates.

In figure 4, we show the results for 4 different DAGs, with multiplicative and additive dynamics (though these are the same for forks and chains). The 6 different dynamics are represented in 4 different DAGs, 2 different correlations, 4 different partial correlations, and 2 different mutual information structures, which means that each of these descriptions is degenerate in some of the dynamics. Partial correlations come close in disentangling direct from indirect effects, but fail to distinguish additive from multiplicative dynamics. Note that we only focus on the sign of the association, and the significance, since the precise value depends on the noise level σ^2. The rightmost column shows that the MFIs assign a unique association structure to each of the dynamics, distinguish between direct and indirect effects, and reveal multiplicative dynamics as a 3-point interaction. Note that both the partial correlation and the MFIs assign a negative association to the parent nodes in a collider structure. This reflects that two nodes become dependent when conditioned on a common effect (*cf.* Berkson's paradox), a spurious effect already found in partial correlations of metabolomic data in [19].

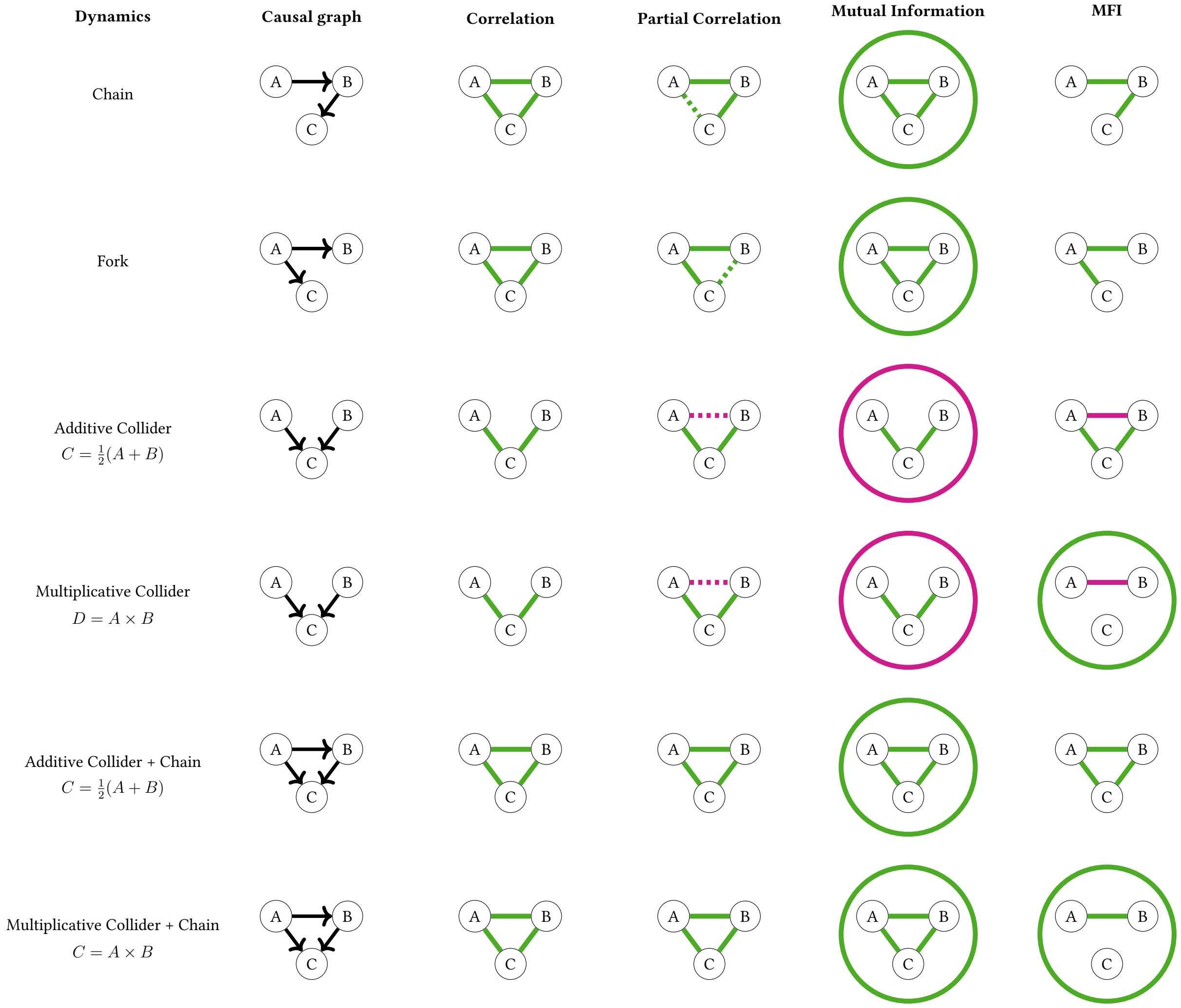


Figure 4: Different causal dynamics lead to different association metrics. Green edges denote positive values, red edges denote negative values, circles denote a 3-point quantity, and dashed lines show edges that show marginal significance, depending on σ^2. Correlations and mutual information cannot distinguish between most dynamics, and while partial correlation can, for certain noise levels, identify the correct pairwise relationships, it falls short of distinguishing additive from multiplicative dynamics. Only MFIs distinguish between all 6 scenarios, and reveal the combinatorial effect of a multiplicative interaction. See appendix A.2 for the simulation parameters and raw numbers.

4.3 Higher-order categorical interactions distinguish dy- and triadic interactions

That the interactions have such resolving power over distributions of binary variables is perhaps not so surprising in light of the universality of RBMs with respect to this class of distributions. More surprisingly, their resolving power extends to the case of categorical variables. In [17], the authors introduce two distributions, the dy- and triadic distributions, that are indistinguishable by all Shannon-like information measures (*e.g.* Shannon-, Renyi(2)-, residual-, and Tsallis entropy, co-information, total correlation, CAEKL mutual information, interaction information, Wyner-, exact-, functional-, and MSS common information, perplexity, disequilibrium, and the LMRP- and TSE complexity).

The two distributions are defined on 3 variables, each taking a value in a 4-letter alphabet $\{0, 1, 2, 3\}$. The joint probabilities are summarised in table 4. To construct the distributions, each category is represented as a binary string ($0 = (00), 3 = (11)$), leading to new variables $\{X_0, X_1, Y_0, Y_1, Z_0, Z_1\}$. The dyadic distribution is constructed by linking these new variables with pairwise rules: $X_0 = Y_1, Y_0 = Z_1, Z_0 = X_1$, while the triadic distribution is constructed with rules involving triplets: $X_0 + Y_0 + Z_0 = 0 \mod 2$, and $X_1 = Y_1 = Z_1$. The resulting binary strings are then reinterpreted as categorical variables to produce table 4.
The authors of [17] find that no Shannon-like measure can distinguish between the two distributions, and they argue that the partial information decomposition, which is different for the two distributions, is not a natural information measure since it has to single out one of the variabels as an output. To calculate model-free categorical interactions between the variables, we set the probabilities of the states in table 4 uniformly to $p = (1 - (64 - 8)\epsilon)/8$, and of the other states to ϵ (so that we have a normalised uniform distribution over legal states). There are a total of $6^3 = 216$ interactions such that $x_1 > x_0, y_1 > y_0, z_1 > z_0$. Each of these can be written as:

$$\begin{aligned}
&I_{XYZ}(x_0 \to x_1; y_0 \to y_1; z_0 \to z_1) = \\
&\log \frac{p\Bigl(X = x_1, Y = y_1, Z = z_1 \mid \underline{X} = 0\Bigr)}{p\Bigl(X = x_0, Y = y_0, , Z = z_0 \mid \underline{X} = 0\Bigr)} \frac{p\Bigl(X = x_1, Y = y_0, Z = z_0 \mid \underline{X} = 0\Bigr)}{p\Bigl(X = x_0, Y = y_1, , Z = z_1 \mid \underline{X} = 0\Bigr)} \\
&\times \frac{p\Bigl(X = x_0, Y = y_1, Z = z_0 \mid \underline{X} = 0\Bigr)}{p\Bigl(X = x_1, Y = y_0, , Z = z_1 \mid \underline{X} = 0\Bigr)} \frac{p\Bigl(X = x_0, Y = y_0, Z = z_1 \mid \underline{X} = 0\Bigr)}{p\Bigl(X = x_1, Y = y_1, , Z = z_0 \mid \underline{X} = 0\Bigr)}
\end{aligned} \tag{78}$$

Of particular interest are two quantities: $I_{XYZ}(0 \to 3; 0 \to 3; 0 \to 3)$, and $\overline{I}_{XYZ} = \sum_{x_0,x_1,y_0,y_1,z_0,z_1} I_{XYZ}(x_0 \to x_1; y_0 \to y_1; z_0 \to z_1)$, where the sum goes over all values such that $x_1 > x_0, y_1 > y_0, z_1 > z_0$, since all possible pairs necessarily sum to zero as $I_{XYZ}(x_0 \to x_1; y_0 \to y_1; z_0 \to z_1) = -I_{XYZ}(x_1 \to x_0; y_0 \to y_1; z_0 \to z_1)$.

For the dyadic distribution, we have

$$I^{\text{Dy}}_{XYZ}(0 \to 3; 0 \to 3; 0 \to 3) = \log \frac{p\epsilon^3}{p\epsilon^3} = 0 \tag{79}$$

While for the triadic distribution:

$$I^{\text{Tri}}_{XYZ}(0 \to 3; 0 \to 3; 0 \to 3) = \log \frac{\epsilon^4}{p\epsilon^3} = \log \frac{\epsilon}{p} \tag{80}$$

So this particular 3-point interaction is zero for the dyadic, and negative for the triadic distribution. The sum over all 3-points is (see appendix A.3 for details):

$$\overline{I}^{\text{Dy}}_{XYZ} = \log 1 = 0 \tag{81}$$

$$\overline{I}^{\text{Tri}}_{XYZ} = 64 \log \frac{\epsilon}{p} \tag{82}$$

That is, the additively symmetrised 3-point interaction is zero for the dyadic distribution, and strongly negative for the triadic distribution. These two distributions that are indistinguishable in terms of information structure are distinguishable by their model-free interactions, and these accurately reflect the higher-order nature of the triadic distribution.

Dyadic

X	Y	Z	P
0	0	0	1 / 8
0	2	1	1 / 8
1	0	2	1 / 8
1	2	3	1 / 8
2	1	0	1 / 8
2	3	1	1 / 8
3	1	2	1 / 8
3	3	3	1 / 8

Triadic

X	Y	Z	P
0	0	0	1 / 8
1	1	1	1 / 8
0	2	2	1 / 8
1	3	3	1 / 8
2	0	2	1 / 8
3	1	3	1 / 8
2	2	0	1 / 8
3	3	1	1 / 8

Table 4: The joint probability of the dy- and triadic distributions (from [17]). All other states have probability zero.

5 Discussion

In this paper, we have related the model-free interactions introduced in [4] to information theory by defining them as Möbius inversions of surprisal on the same lattice that relates mutual information to entropy. We then inverted the order of the lattice and computed the order-dual to the mutual information, which turned out to be a generalisation of differential mutual information. Similarly the order-dual of interaction turned out to be interaction in a different context. Both the in- and outeractions are able to distinguish all six logic gates by value and sign. Moreover, their absolute strength reflects the synergy within the logic gate. On simulations, the interactions were able to perfectly distinguish six kinds of causal dynamics that were partially indistinguishable to Pearson/partial correlations, causal graphs, and mutual information. Finally, we considered the dy- and triadic distributions, which are constructed using pairwise, or higher-order rules, respectively. These two distributions are indistinguishable in terms of their Shannon-information, but have different categorical MFIs, which reflect the order of the construction rules.

The interactions gain this advantage over entropy-based quantities by being defined in a pointwise way, *i.e.* in terms of the surprisal of particular states, rather than their expectation values across all states. It was already known that the MFIs are equivalent to Ising interactions, which in turn are equivalent to the effective couplings in Restricted Boltzmann machines. As restricted Boltzmann machines are known to be universal approximators, the MFIs should in theory be able to perfectly characterise any distribution. One might worry that precise probabilities are harder to estimate than expectation values, but they can be directly estimated from an empirical sample.

A major limitation of the MFIs is that they are only defined on binary or categorical variables, whereas many of the other association metrics are defined for ordinal and continuous variables as well. States of continuous variables no longer form a lattice, so it is hard to see how one could extend the definition of MFIs to include these cases.

Another thing the attentive reader might worry about is the stringency in the conditioning. Estimating $\log p(X = 1, Y = 1, T = 0)$ directly from data means counting states that look like $(X, Y, T_1, T_2, \dots, T_N) = (1, 1, 0, 0, \dots 0)$, which for sufficiently large N will be rare in any data set. In an upcoming paper, we address this by estimating the graph of conditional dependencies, and only conditioning on the Markov blanket of each node.

Finally, it is worth noting that the structure of different lattices has guided much of this research. That Boolean algebras are important in defining higher-order structure is not so surprising, since they are the stage on which one can generalise the inclusion-exclusion principle [29], but not only does their order-reversed dual lead to meaningful definitions, so do completely unrelated lattices. For example, the Möbius inversion on the lattice of ordinal variables from figure 3, or the redundancy lattices in the partial information decomposition [36], both lead to new, sensible definitions of information theoretic quantities. Furthermore, the notion of Möbius inversion has been generalised to categories [21], of which posets are a special case. A systematic investigation of information-theoretical quantities in this richer context would be most interesting.

Acknowledgements

The author is thankful to Ava Khamseh, Sjoerd Beentjes, Chris Ponting, and Luigi Del Debbio for helpful discussions, and to Sjoerd Beentjes for reading and commenting on an early version of this manuscript. AJ is supported by the University of Edinburgh and University of Glasgow joint Precision Medicine DTP.

References

[1] U. Alvarez-Rodriguez, F. Battiston, G. F. de Arruda, Y. Moreno, M. Perc, and V. Latora. Evolutionary dynamics of higher-order interactions in social networks. *Nature Human Behaviour*, 5(5):586–595, 2021.

[2] F. Bartolucci, R. Colombi, and A. Forcina. An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica*, pages 691–711, 2007.

[3] G. Bateson. *Steps to an Ecology of Mind*. Chandler Publishing Company, 1972.

[4] S. V. Beentjes and A. Khamseh. Higher-order interactions in statistical physics and machine learning: A model-independent solution to the inverse problem at equilibrium. *Phys. Rev. E*, 102:053314, Nov 2020. doi: 10.1103/PhysRevE.102.053314. URL https://link.aps.org/doi/10.1103/PhysRevE.102.053314.

[5] A. J. Bell. The co-information lattice. In *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA*, volume 2003. Citeseer, 2003.

[6] G. Cencetti, F. Battiston, B. Lepri, and M. Karsai. Temporal properties of higher-order interactions in social networks. *Scientific reports*, 11(1):1–10, 2021.

[7] N. J. Cerf and C. Adami. Entropic bell inequalities. *Physical Review A*, 55(5):3371, 1997.

[8] G. Cossu, L. Del Debbio, T. Giani, A. Khamseh, and M. Wilson. Machine learning determination of dynamical parameters: The ising model case. *Physical Review B*, 100(6):064304, 2019.

[9] Y. Freund and D. Haussler. Unsupervised learning of distributions of binary vectors using two layer networks. 1994.

[10] D. J. Galas, N. A. Sakhanenko, A. Skupin, and T. Ignac. Describing the complexity of systems: Multivariable "set complexity" and the information basis of systems biology. *Journal of Computational Biology*, 21(2):118–140, 2014.

[11] D. J. Galas, J. Kunert-Graf, L. Uechi, and N. A. Sakhanenko. Towards an information theory of quantitative genetics, 2019.

[12] E. Ganmor, R. Segev, and E. Schneidman. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proceedings of the National Academy of Sciences*, 108(23):9679–9684, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1019641108. URL https://www.pnas.org/content/108/23/9679.

[13] M. Gatica, R. Cofré, P. A. Mediano, F. E. Rosas, P. Orio, I. Diez, S. P. Swinnen, and J. M. Cortes. High-order interdependencies in the aging brain. *Brain connectivity*, 11 (9):734–744, 2021.

[14] G. F. Glonek and P. McCullagh. Multivariate logistic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):533–546, 1995.

[15] M. Grabisch and M. Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28(4): 547–565, 1999.

[16] J. Grilli, G. Barabás, M. J. Michalska-Smith, and S. Allesina. Higher-order interactions stabilize dynamics in competitive network models. *Nature*, 548(7666):210–213, 2017.

[17] R. G. James and J. P. Crutchfield. Multivariate dependence beyond shannon information. *Entropy*, 19(10):531, 2017.

[18] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 1957. ISSN 0031899X. doi: 10.1103/PhysRev.106.620.

[19] J. Krumsiek, K. Suhre, T. Illig, J. Adamski, and F. J. Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology*, 5(1):21, 2011.

[20] E. Kuzmin, B. VanderSluis, W. Wang, G. Tan, R. Deshpande, Y. Chen, M. Usaj, A. Balint, M. M. Usaj, J. Van Leeuwen, E. N. Koch, C. Pons, A. J. Dagilis, M. Pryszlak, J. Z. Y. Wang, J. Hanchard, M. Riggi, K. Xu, H. Heydari, B. J. S. Luis, E. Shuteriqi, H. Zhu, N. Van Dyk, S. Sharifpoor, M. Costanzo, R. Loewith, A. Caudy, D. Bolnick, G. W. Brown, B. J. Andrews, C. Boone, and C. L. Myers. Systematic analysis of complex genetic interactions. *Science*, 2018. ISSN 10959203. doi: 10.1126/science.aao1729.

[21] T. Leinster. Notions of möbius inversion. *Bulletin of the Belgian Mathematical Society-Simon Stevin*, 19(5):909–933, 2012.

[22] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 2006. ISSN 00278424. doi: 10.1073/pnas.0609152103.

[23] Y. Li, M. M. Mayfield, B. Wang, J. Xiao, K. Kral, D. Janik, J. Holik, and C. Chu. Beyond direct neighbourhood effects: higher-order interactions improve modelling and predicting tree survival and growth. *National Science Review*, 8(5):nwaa244, 2021.

[24] H. Matsuda. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Physical review E*, 62(3):3096, 2000.

[25] H. C. Nguyen, R. Zecchina, and J. Berg. Inverse statistical problems: from the inverse Ising problem to data science. *Advances in Physics*, 2017. ISSN 14606976. doi: 10.1080/00018732.2017.1341604.

[26] D. Panas, A. Maccione, L. Berdondini, and M. H. Hennig. Homeostasis in large networks of neurons through the Ising model - do higher order interactions matter? *BMC Neuroscience*, 2013. ISSN 1471-2202. doi: 10.1186/1471-2202-14-s1-p166.

[27] G.-C. Rota. On the foundations of combinatorial theory i. theory of möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 2(4):340–368, 1964.

[28] A. Sanchez. Defining Higher-Order Interactions in Synthetic Ecology: Lessons from Physics and Quantitative Genetics. *Cell Systems*, 2019. ISSN 24054720. doi: 10.1016/j.cels.2019.11.009.

[29] R. P. Stanley. Enumerative combinatorics volume 1 second edition. *Cambridge studies in advanced mathematics*, 2011.

[30] E. Tekin, C. White, T. M. Kang, N. Singh, M. Cruz-Loya, R. Damoiseaux, V. M. Savage, and P. J. Yeh. Prevalence and patterns of higher-order drug interactions in escherichia coli. *NPJ systems biology and applications*, 4(1):1–10, 2018.

[31] G. Tkačik, O. Marre, D. Amodei, E. Schneidman, W. Bialek, and M. J. Berry. Searching for Collective Behavior in a Large Network of Sensory Neurons. *PLoS Computational Biology*, 2014. ISSN 1553734X. doi: 10.1371/journal.pcbi.1003408.

[32] S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

[33] J. Watkinson, K. C. Liang, X. Wang, T. Zheng, and D. Anastassiou. Inference of regulatory gene interactions from expression data using three-way mutual information. *Annals of the New York Academy of Sciences*, 2009. ISSN 17496632. doi: 10.1111/j.1749-6632.2008.03757.x.

[34] D. M. Weinreich, Y. Lan, C. S. Wylie, and R. B. Heckendorn. Should evolutionary geneticists worry about higher-order epistasis? *Current opinion in genetics & development*, 23(6):700–707, 2013.

[35] M. Wibral, V. Priesemann, J. W. Kay, J. T. Lizier, and W. A. Phillips. Partial information decomposition as a unified approach to the specification of neural goal functions. *Brain and cognition*, 112:25–38, 2017.

[36] P. L. Williams and R. D. Beer. Nonnegative Decomposition of Multivariate Information. pages 1–14, 2010. URL http://arxiv.org/abs/1004.2515.

[37] S. Yu, H. Yang, H. Nakahara, G. S. Santos, D. Nikolić, and D. Plenz. Higher-order interactions characterized in cortical activity. *Journal of neuroscience*, 31(48):17514–17526, 2011.

A Appendix

A.1 Proofs

Estimating the interaction in Definition 2 from data involves estimating probabilities $P(X)$ of certain states X occurring. We do not have access to the true probabilities, but can rewrite the interactions in terms of expectation values. Note that all interactions involve factors of the type

$$\frac{p(X=1, Y=y \mid Z=0)}{p(X=0, Y=y \mid Z=0)} = \frac{p(X=1 \mid Y=y, Z=0)}{p(X=0 \mid Y=y, Z=0)} \tag{83}$$

$$= \frac{p(X=1 \mid Y=y, Z=0)}{1-p(X=1 \mid Y=y, Z=0)} \tag{84}$$

Which can be written as

$$= \frac{\mathbb{E}[X \mid Y=y, Z=0]}{1-\mathbb{E}[X \mid Y=y, Z=0]} \tag{85}$$

since

$$\mathbb{E}[X \mid Z=z] = \sum_{x \in \{0,1\}} p(X=x \mid Z=z)\, x = p(X=1 \mid Z=z) \tag{86}$$

Which allows us to write *e.g.* the 2-point interaction as

$$I_{ij} = \log \frac{\mathbb{E}\left(X_i | X_j=1, \underline{X}=0\right)\left(1-\mathbb{E}\left(X_i | X_j=0, \underline{X}=0\right)\right)}{\mathbb{E}\left(X_i | X_j=0, \underline{X}=0\right)\left(1-\mathbb{E}\left(X_i | X_j=1, \underline{X}=0\right)\right)} \tag{87}$$

Expectation values are theoretical, not empirical, quantities, but one can use sample means as unbiased estimators to estimate each term in (87). The stringent conditioning in this estimator can make the number of samples that satisfy the conditioning very small, which gives the estimates a large variance on different finite samples. Note that if we can find a subset of variables MB_{X_i} such that $X_i \perp\!\!\!\perp X_k \mid \text{MB}_{X_i} \;\; \forall X_k \notin \text{MB}_{X_i}$ and $i \neq k$ (in causal language: a set of variables MB_{X_i} that d-separates X_i from the rest), then we only have to condition on MB_{X_i} in (87), reducing the variance of our estimator. Such a set MB_{X_i} is called a *Markov blanket*[5] of the node X_i. Since conditioning on fewer variables should reduce the variance of the estimate by increasing the number of samples that can be used for the estimation, we are generally interested in finding the smallest Markov blanket. This minimal Markov blanket is also called the Markov boundary.

Finding such Markov blankets is hard. I fact, since it requires testing each possible conditional dependency between the variables, I claim (without proof) it is *causal-discovery*-hard, *i.e.* it is at least as computationally complex as constructing a causal DAG consistent with the joint probability distribution, if such a graph exists.

The Markov blankets are not only a computational trick – in theory, only variables that are in each other's Markov blanket can share a nonzero interaction. To see this, first note that the property of being in a variable's Markov blanket is symmetric:

Proposition 1. *(Symmetry of Markov blankets) Let X be a set of variables with joint distribution $p(X)$. Let $A \in X$ and $B \in X$ such that $A \neq B$. Denote the minimal Markov blanket of X by MB_X. Then $A \in MB_B \iff B \in MB_A$, and we say that A and B are Markov-connected.*

[5] There has recently been some confusion on the notion of Markov blankets in biology, specifically w.r.t. their use in the free energy principle of neuroscience. When I say Markov blanket, I am referring to the notion of a *Pearl blanket* in the language of [**?**].

Proof. Let $Y = X \setminus \{A, B\}$. Then

$$A \notin \text{MB}_B \implies p(B \mid A, Y) = p(B \mid Y) \tag{88}$$

Consider

$$\begin{aligned} p(A \mid B, Y) &= \frac{p(A, B \mid Y)}{p(B \mid X)} && (89) \\ &= \frac{p(B \mid A, Y) p(A, \mid Y)}{p(B \mid Y)} && (90) \\ &= p(A \mid Y) && (91) \end{aligned}$$

Which means that $B \notin \text{MB}_A$. Since $A \notin \text{MB}_B \iff B \notin \text{MB}_A$ holds, its negation also holds, which completes the proof. □

This definition of Markov-connectedness allows us to state the following:

Theorem 2. *(Only Markov-connected variables interact) A model-free n-point interaction $I_{1\dots n}$ can only be nonzero when all variables $S = \{X_1, \dots, X_n\}$ are mutually Markov-connected.*

Proof. Let X be a set of variables with joint distribution $p(X)$. Let $S = \{X_1, \dots, X_n\}$, and $\underline{X} = X \setminus S$. Consider the definition of an n-point interaction among S:

$$\begin{aligned} I_{1\dots n} &= \prod_{i=1}^{n} \frac{\partial}{\partial X_i} \log p(X_1, \dots, X_n \mid \underline{X}) && (92) \\ &= \left(\prod_{i=1}^{n-1} \frac{\partial}{\partial X_i} \right) \frac{\partial}{\partial X_n} \log p(X_1, \dots, X_n \mid \underline{X}) && (93) \\ &= \left(\prod_{i=1}^{n-1} \frac{\partial}{\partial X_i} \right) \log \frac{p(X_n = 1 \mid X_1, \dots, X_{n-1}, \underline{X})}{p(X_n = 0 \mid X_1, \dots, X_{n-1}, \underline{X})} && (94) \\ &= \left(\prod_{i=1}^{n-1} \frac{\partial}{\partial X_i} \right) \log \frac{p(X_n = 1 \mid S \setminus X_n, \underline{X})}{p(X_n = 0 \mid S \setminus X_n, \underline{X})} && (95) \end{aligned}$$

Now, if $\exists X_j \in S$ such that $X_j \notin \text{MB}_{X_n}$, then we do not need to condition on X_j and can write this as

$$\begin{aligned} &= \left(\prod_{i=1}^{n-1} \frac{\partial}{\partial X_i} \right) \log \frac{p(X_n = 1 \mid S \setminus \{X_j, X_n\}, \underline{X})}{p(X_n = 0 \mid S \setminus \{X_j, X_n\}, \underline{X})} && (96) \\ &= \left(\prod_{\substack{i=1 \\ i \neq j}}^{n-1} \frac{\partial}{\partial X_i} \right) \left(\frac{\partial}{\partial X_j} \log \frac{p(X_n = 1 \mid S \setminus \{X_j, X_n\}, \underline{X})}{p(X_n = 0 \mid S \setminus \{X_j, X_n\}, \underline{X})} \right) && (97) \\ &= 0 && (98) \end{aligned}$$

since the probabilities no longer involve X_j. Since X_j was chosen without loss of generality, this must hold for all variables in S, which means that if any variable in S is not in the Markov blanket of X_n, then the interaction I_S vanishes:

$$S \setminus X_n \not\subset \text{MB}_{X_n} \implies I_S = 0 \tag{99}$$

Furthermore, the indexing we chose for our variables was arbitrary, so this must hold for any reindexing, which means that

$$\forall X_i \in S: \quad S \setminus X_i \not\subset \text{MB}_{X_i} \implies I_S = 0 \tag{100}$$

Which means that all variables in S must be Markov-connected for the interaction I_S to be nonzero. □

Knowledge of the causal graph thus helps estimation in two ways: it shrinks the variance of the estimates by relaxing the conditioning, and in addition identifies the interactions that could be nonzero.

If your knowledge of the causal graph in imperfect, you might accidentally exclude a variable from a Markov blanket and undercondition the relevant probabilities. The error that this results in can be expressed in terms of the mutual information between the variables:

Proposition 2. *(Underconditioning bias) Let S be a set of random variables with probability distribution $p(S)$. Let X, Y, and Z be three disjoint subsets of S. Then omitting Y from the conditioning set results in a bias determined by, and linear in, the pointwise mutual information that $Y = 0$ gives about states of X:*

$$I_{X|YZ} - I_{X|Z} = \left(\prod_{i=1}^{|X|} \frac{\partial}{\partial x_i} \right) \text{pmi}(X = x, Y = 0 \mid Z = 0) \tag{101}$$

Proof. The pointwise mutual information (pmi) is defined as

$$\text{pmi}(X = x, Y = y) = \log \frac{p(X = x, Y = y)}{p(X = x)p(Y = y)} \tag{102}$$

Note that

$$p(X = x_1 \mid Y = y, Z = z) = \frac{p(X = x_1, Y = y \mid Z = z)}{p(Y = y \mid Z = z)} \tag{103}$$

So that we can write

$$p(X = x_1 \mid Y = y, Z = z) = e^{\text{pmi}(X=x_1,Y=y|Z=z)} p(X = x_1 \mid Z = z) \tag{104}$$

That is, not conditioning on $Y = y$ results in an error in the estimate of $p(X = x_1 \mid Y = y, Z = z)$ that is exponential in the Z-conditional pmi of X and Y. However, consider the interaction among X:

$$I_X = I_{X|YZ} = \left(\prod_{i=1}^{|X|} \frac{\partial}{\partial x_i} \right) \log p(X = x \mid Y = 0, Z = 0) \tag{105}$$

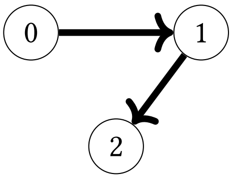
$$= \left(\prod_{i=1}^{|X|} \frac{\partial}{\partial x_i} \right) (\log p(X = x \mid Z = 0) + \text{pmi}(X = x, Y = 0 \mid Z = 0)) \tag{106}$$

$$= I_{X|Z} + \left(\prod_{i=1}^{|X|} \frac{\partial}{\partial x_i} \right) \text{pmi}(X = x, Y = 0 \mid Z = 0) \tag{107}$$

That is, the error in the interaction as a result of not conditioning on the right variables is linear in the difference between the pmi's of different states. □

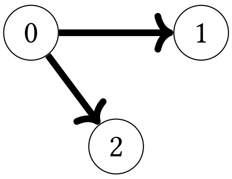
A.2 Numerics of causal structures

In Tables 5 to 10, we list the precise numbers that led to figure 4. From each graph, we generate 100k samples using $p = 0.5$ and $\sigma = 0.4$. To quantify the significance value of the interactions, we bootstrap resample the data 1k times, and calculate F: the fraction of resampled interactions that have a different sign from the original interaction. The smaller F is, the more significant the interaction is.



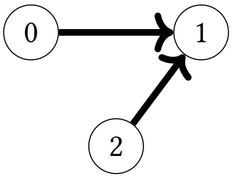
	Genes	Interaction	F	Pearson cor.	Pearson cor. p	Partial cor.	Partial cor. p	MI
0	[0, 1]	4.281	0.000	0.790	0.0	0.635	0.000e+00	0.515
1	[0, 2]	0.056	0.117	0.622	0.0	0.031	2.261e-23	0.301
2	[1, 2]	4.249	0.000	0.786	0.0	0.628	0.000e+00	0.510
3	[0, 1, 2]	-0.052	0.217	NaN	NaN	NaN	NaN	0.300

Table 5: Chain



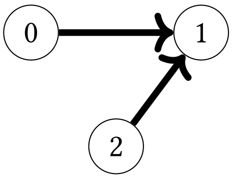
	Genes	Interaction	F	Pearson cor.	Pearson cor. p	Partial cor.	Partial cor. p	MI
0	[0, 1]	4.268	0.000	0.789	0.0	0.634	0.000e+00	0.514
1	[0, 2]	4.257	0.000	0.788	0.0	0.632	0.000e+00	0.512
2	[1, 2]	-0.014	0.376	0.622	0.0	0.028	6.518e-19	0.300
3	[0, 1, 2]	0.020	0.376	NaN	NaN	NaN	NaN	0.300

Table 6: Fork



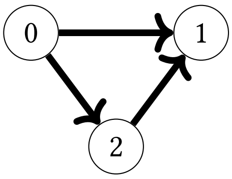
	Genes	Interaction	F	Pearson cor.	Pearson cor. p	Partial cor.	Partial cor. p	MI
0	[0, 1]	2.144	0.000	0.395	0.000	0.505	0.000e+00	1.154e-01
1	[0, 2]	-0.989	0.000	-0.002	0.593	-0.070	5.172e-109	2.059e-06
2	[1, 2]	2.144	0.000	0.395	0.000	0.505	0.000e+00	1.154e-01
3	[0, 1, 2]	0.003	0.438	NaN	NaN	NaN	NaN	-2.678e-02

Table 7: Additive collider



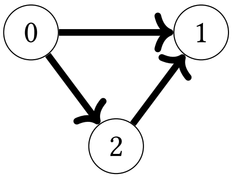
	Genes	Interaction	F	Pearson cor.	Pearson cor. p	Partial cor.	Partial cor. p	MI
0	[0, 1]	0.032	0.140	0.427	0.000	0.478	0.000e+00	1.403e-01
1	[0, 2]	-2.156	0.000	-0.005	0.145	-0.087	1.463e-166	1.529e-05
2	[1, 2]	0.036	0.109	0.429	0.000	0.480	0.000e+00	1.415e-01
3	[0, 1, 2]	4.237	0.000	NaN	NaN	NaN	NaN	-1.150e-01

Table 8: Multiplicative collider



	Genes	Interaction	F	Pearson cor.	Pearson cor. p	Partial cor.	Partial cor. p	MI
0	[0, 1]	2.103	0.000	0.705	0.0	0.362	0.0	0.396
1	[0, 2]	3.288	0.000	0.790	0.0	0.599	0.0	0.515
2	[1, 2]	2.113	0.000	0.706	0.0	0.364	0.0	0.397
3	[0, 1, 2]	0.050	0.162	NaN	NaN	NaN	NaN	0.335

Table 9: Additive collider + chain



	Genes	Interaction	F	Pearson cor.	Pearson cor. p	Partial cor.	Partial cor. p	MI
0	[0, 1]	-0.017	0.342	0.709	0.0	0.365	0.0	0.403
1	[0, 2]	2.094	0.000	0.786	0.0	0.596	0.0	0.510
2	[1, 2]	-0.057	0.092	0.707	0.0	0.361	0.0	0.401
3	[0, 1, 2]	4.359	0.000	NaN	NaN	NaN	NaN	0.293

Table 10: Multiplicative collider + chain

A.3 Python code to calculate categorical dy- and triadic interactions

```
dyadicStates = [['a', 'a', 'a'], ['a', 'c', 'b'], ['b', 'a', 'c'], ['b', 'c', 'd'],
['c', 'b', 'a'], ['c', 'd', 'b'], ['d', 'b', 'c'], ['d', 'd', 'd']]

triadicStates = [['a', 'a', 'a'], ['a', 'c', 'c'], ['b', 'b', 'b'], ['b', 'd', 'd'],
['c', 'a', 'c'], ['c', 'c', 'a'], ['d', 'b', 'd'], ['d', 'd', 'b']]

stateDict = {0: 'a', 1: 'b', 2:'c', 3: 'd'}

def catIntSymb(x0, x1, y0, y1, z0, z1, states):
    prob = lambda x, y, z: 'p' if [x, y, z] in states else 'e'

    num = prob(x1, y1, z1) + prob(x1, y0, z0) + prob(x0, y1, z0) + prob(x0, y0, z1)
    denom = prob(x1, y1, z0) + prob(x1, y0, z1) + prob(x0, y1, z1) + prob(x0, y0, z0)
    return (num, denom)

numDy = ''
denomDy = ''
numTri = ''
denomTri = ''

for x0 in range(4):
    for x1 in range(x0+1, 4):
        for y0 in range(4):
            for y1 in range(y0+1, 4):
                for z0 in range(4):
                    for z1 in range(z0+1, 4):

                        nDy, dDy = catIntSymb(*[stateDict[x] for x in [x0, x1, y0, y1, z0, z1]], dyadicStates)
                        numDy += nDy
                        denomDy += dDy

                        nTri, dTri = catIntSymb(*[stateDict[x] for x in [x0, x1, y0, y1, z0, z1]], triadicStates)
                        numTri += nTri
                        denomTri += dTri


print(f'Dyadic␣interaction:␣log␣(p^{numDy.count("p")␣-␣denomDy.count("p")}␣e^{numDy.count("e")␣-␣denomDy.count("e")})')
print(f'Triadic␣interaction:␣log␣(p^{numTri.count("p")␣-␣denomTri.count("p")}␣e^{numTri.count("e")␣-␣denomTri.count("e")})')

// Output:

>> Dyadic interaction: log (p^0 e^0)
>> Triadic interaction: log (p^-64 e^64)
```