Unique Information via Dependency Constraints

Ryan G. James
Jeffrey Emenheiser
James P. Crutchfield

SFI WORKING PAPER: 2017-09-034

SANTA FE INSTITUTE

Santa Fe Institute Working Paper 17-09-XXX
arXiv:1709.XXXX

Unique Information via Dependency Constraints

Ryan G. James,* Jeffrey Emenheiser,† and James P. Crutchfield‡
*Complexity Sciences Center and Physics Department,
University of California at Davis, One Shields Avenue, Davis, CA 95616*

(Dated: September 19, 2017)

The partial information decomposition is perhaps the leading proposal for resolving shared information in a joint random variable into redundant, synergistic, and unique constituents. Unfortunately, the framework has been hindered by a lack of a generally agreed-upon, multivariate method of quantifying the constituents. Here, we take a step toward rectifying this by developing a decomposition based on a new method that quantifies unique information. The result is the first measure which satisfies the core axioms of the framework while also not treating identical but independent channels as redundant. This marks a key step forward in the practical application of the partial information decomposition.

PACS numbers: 05.45.-a 89.75.Kd 89.70.+c 02.50.-r
Keywords: partial information decomposition, mutual information, statistical dependence, information theory, cybernetics.

I. INTRODUCTION

Understanding how information is stored, modified, and transmitted among the components of a complex system is fundamental to the sciences. Application domains where this would be particularly enlightening include gene regulatory networks, neural coding, highly-correlated electron systems, financial markets, and other complex systems whose large-scale organization is either not known a priori or emerges spontaneously.

One particularly promising framework for accomplishing this is the *partial information decomposition* [1]. Once a practitioner partitions a given set of random variables into *sources* and a *target*, the framework decomposes the information shared between the two sets into interpretable, nonnegative components—in the bivariate case: redundant, unique, and synergistic informations. This task relies on two separate aspects of the framework: first, the overlapping source subsets (algebraic lattice [2]) into which the information should be decomposed and, second, the method of quantifying those components. Unfortunately, despite a great deal of effort [3–12], the current consensus is that the lattice needs to be modified [9–11, 13, 14] and that extant methods of quantifying components [1, 3–5, 7, 8] are not satisfactory in full multivariate generality. Thus, the promise of a full informational analysis of the organization of complex systems remains unrealized.

The following addresses the second aspect—quantifying the components. Inspired by early cybernetics—specifically, Ref. [15]'s lattice of system models—we develop a general technique for decomposing arbitrary multivariate information measures according to how they are influenced by statistical dependencies. We then utilize this decomposition to quantify the information that one variable uniquely has about another. Reference [5]'s $\mathrm{I}_{\mathrm{BROJA}}$ measure also quantifies unique information, but is susceptible to artificially inflating redundancy [8]. Both our measure as well as Ref. [8]'s $\mathrm{I}_{\mathrm{ccs}}$ measure overcome this issue, but $\mathrm{I}_{\mathrm{ccs}}$ does so at the expense of positivity. This makes our proposal the only method of quantifying the partial information decomposition that is both nonnegative and does not artificially inflate redundancy.

Our development proceeds as follows. Section II reviews the partial information decomposition and Section III introduces our measure of unique information. Section IV then compares our measure to others on a variety of exemplar distributions, exploring and contrasting its behavior. Section V discusses several open conceptual issues and Section VI concludes. The development requires a working knowledge of information theory, such as found in standard texts [16–18].

II. BACKGROUND

Consider a set of *sources* $X_0, X_1, \ldots, X_{n-1} = X_{0:n}$ and a *target* Y.[1] The amount of information the sources carry about the target is quantified by their mutual information:

$$\mathrm{I}\left[X_0, X_1, \ldots, X_{n-1} : Y\right] = \mathrm{I}\left[X_{0:n} : Y\right] \ .$$

* rgjames@ucdavis.edu
† jemenheiser@ucdavis.edu
‡ chaos@ucdavis.edu

[1] We subscript the joint variable with a Python-like array-slice notation.

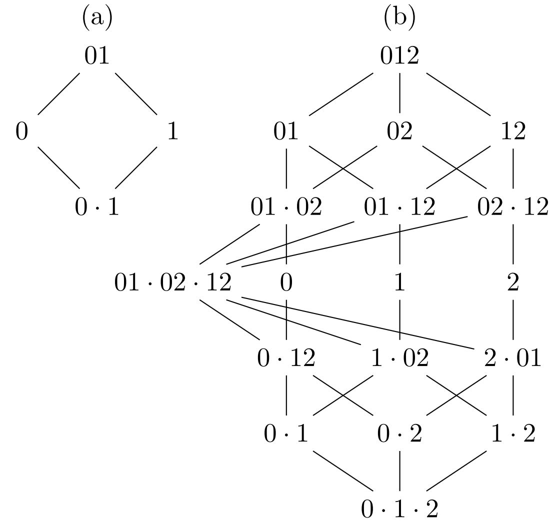


FIG. 1. Lattice of antichains for (a) two (X_0 and X_1) and (b) three sources (X_0, X_1, and X_2): An antichain is represented using a dot to separate sets and sets by concatenated indices; *e.g.*, $\{\{X_0\}\{X_1, X_2\}\}$ is represented $0 \cdot 12$.

The partial information decomposition (PID) [1] then assigns *shared information* to sets of source groupings such that no (inner) set is subsumed by another. In this way, the PID quantifies what information about the target each of those groups has in common.

A. Antichain Lattices

The sets of groupings we consider are *antichains*:

$$\mathcal{A}(X_{0:n}) = \left\{\alpha \in \mathcal{P}^+\big(\mathcal{P}^+(X_{0:n})\big) : \forall s_1, s_2 \in \alpha, s_1 \not\subset s_2\right\}$$

where $\mathcal{P}^+(S) = \mathcal{P}(S) \setminus \{\emptyset\}$ denotes the set of nonempty subsets of set S. Antichains form a algebraic lattice [2], where one antichain α is less than another β if the information shared among the groupings of variables in the first is included in the shared information of the second:

$$\alpha \preceq \beta \iff \forall s_1 \in \beta, \exists s_2 \in \alpha, s_2 \subseteq s_1 \ . \tag{1}$$

Figure 1 graphically depicts antichain lattices for two and three variables. There, for brevity's sake, a dot separates the sets within an antichain, and the groups of sources are represented by their indices concatenated. For example, $0 \cdot 12$ represents the antichain $\{\{X_0\}\{X_1, X_2\}\}$.

B. Shared Informations

Given the antichain lattice, one then assigns a quantity of shared or redundant information to each. This should quantify the amount of information shared by each set of sources within an antichain α about the target. This shared information will be denoted $\mathrm{I}_\cap[\alpha \to Y]$. Reference [1] put forth several axioms that such a measure should follow:

(S) $\mathrm{I}_\cap[\alpha \to Y]$ is unchanged under permutations of α. *(symmetry)*

(SR) $\mathrm{I}_\cap[i \to Y] = \mathrm{I}[X_i : Y]$. *(self-redundancy)*

(M) For all $\alpha \preceq \beta$, $\mathrm{I}_\cap[\alpha \to Y] \leq \mathrm{I}_\cap[\beta \to Y]$. *(monotonicity)*

With a lattice of shared informations in hand, the *partial information* $\mathrm{I}_\partial[\alpha \to Y]$ is defined as the Möbius inversion [2] of the shared information:

$$\mathrm{I}_\cap[\alpha \to Y] = \sum_{\beta \preceq \alpha} \mathrm{I}_\partial[\beta \to Y] \ . \tag{2}$$

We further require that the following axiom hold:

(LP) $\mathrm{I}_\partial[\alpha \to Y] \geq 0$. *(local positivity)*

This ensures that the partial information decomposition forms a partition of the sources-target mutual information and contributes to the interpretability of the decomposition.

C. The Bivariate Case

In the case of bivariate inputs, the partial information decomposition takes a particularly intuitive form. First, following the self-redundancy axiom **(SR)**, the sources-target mutual information decomposes into four components:

$$\begin{aligned} \mathrm{I}[X_0 X_1 : Y] = \mathrm{I}_\partial[0 \cdot 1 \to Y] + \mathrm{I}_\partial[0 \to Y] \\ + \mathrm{I}_\partial[1 \to Y] + \mathrm{I}_\partial[01 \to Y] \ , \end{aligned} \tag{3}$$

and, again following **(SR)**, each source-target mutual information consists of two components:

$$\mathrm{I}[X_0 : Y] = \mathrm{I}_\partial[0 \cdot 1 \to Y] + \mathrm{I}_\partial[0 \to Y] \tag{4}$$

$$\mathrm{I}[X_1 : Y] = \mathrm{I}_\partial[0 \cdot 1 \to Y] + \mathrm{I}_\partial[1 \to Y] \ . \tag{5}$$

The components have quite natural interpretations. $\mathrm{I}_\partial[0 \cdot 1 \to Y]$ is the amount of information that the two sources X_0 and X_1 *redundantly* carry about the target Y. $\mathrm{I}_\partial[0 \to Y]$ and $\mathrm{I}_\partial[1 \to Y]$ quantify the amount of information that sources X_0 and X_1, respectively, carry *uniquely* about the target Y. Finally, $\mathrm{I}_\partial[01 \to Y]$ is the amount of information that sources X_0 and X_1 *synergistically* or collectively carry about the target Y.

Combining the above decompositions, we see that the operational result of conditioning removes redundancy but expresses synergistic effects:

$$\begin{aligned} \mathrm{I}\left[X_0 : Y | X_1\right] &= \mathrm{I}\left[X_0 X_1 : Y\right] - \mathrm{I}\left[X_1 : Y\right] \\ &= \mathrm{I}_\partial\left[0 \to Y\right] + \mathrm{I}_\partial\left[01 \to Y\right] ~. \end{aligned} \qquad (6)$$

Furthermore, the co-information [19] can be expressed as:

$$\begin{aligned} \mathrm{I}\left[X_0 : X_1 : Y\right] &= \mathrm{I}\left[X_0 : Y\right] - \mathrm{I}\left[X_0 : Y | X_1\right] \\ &= \mathrm{I}_\partial\left[0 \cdot 1 \to Y\right] - \mathrm{I}_\partial\left[01 \to Y\right] ~. \end{aligned} \qquad (7)$$

This illustrates one of the strengths of the partial information decomposition. It explains, in a natural fashion, why the co-information can be negative. It is the difference between a distribution's redundancy and synergy.

The bivariate decomposition's four terms are constrained by the three self-redundancy constraints Eqs. (3) to (5), leaving one degree of freedom: $\mathrm{I}_\partial\left[0 \cdot 1 \to Y\right]$. Therefore, specifying any component of the partial information lattice determines the entire decomposition. In the multivariate case, however, only specifying how to compute $\mathrm{I}_\cap$ values completes the decomposition.

Finally, in the bivariate case one further axiom is employed:

(Id) $\mathrm{I}_\cap\left[0 \cdot 1 \to X_0 X_1\right] = \mathrm{I}\left[X_0 : X_1\right]$ *(identity)*

This axiom ensures that simply concatenating independent inputs does not result in redundant information. Unfortunately, it is not clear how to extend this axiom to the multivariate case or even if it should be extended.

D. Extant Methods

We next describe the four primary existing methods for quantifying the partial information decomposition. There exist other methods ($\mathrm{I}_{\mathrm{mmi}}$ [9], $\mathrm{I}_\wedge$ [7], and $\mathrm{I}_\downarrow$ [4, 9]), though they either suffer from inconsistencies or otherwise have not been adopted.

$\mathrm{I}_{\mathrm{min}}$, the first measure proposed [1], quantifies the average least information the individual sources have about each target value. It has been criticized [3, 4] for its behavior in certain situations. For example, when the target simply concatenates two independent sources, it decomposes those two bits into one bit of redundancy and one of synergy. This is in stark contrast to the more intuitive view that the target contains two bits of unique information—one from each source.

$\mathrm{I}_{\mathrm{proj}}$ quantifies shared information using information geometry [3]. Due to its foundation relying on the Kullback-Leibler divergence, however, it does not have any obvious extension to measuring the shared information in antichains of size three or greater.

As in our approach, $\mathrm{I}_{\mathrm{BROJA}}$ attempts to quantify unique information [5]. It does this by finding the minimum $\mathrm{I}\left[X_i : Y | X_{0:n \setminus i}\right]$ over all distributions that preserve source-target marginal distributions. (The random variable set $X_{0:n\setminus i}$, excludes variable i.) However, due to its decision-theoretic underpinnings, there exist distributions for which its optimization artificially correlates the sources [8]. This leads the measure to quantify identical, though independent, source-target channels as fully redundant. Furthermore, as a measure of unique information, it cannot completely quantify the partial information lattice when the number of sources exceeds two.

Finally, $\mathrm{I}_{\mathrm{ccs}}$ quantifies redundant information by aggregating the pointwise coinformation terms whose signs agree with the signs of all the source-target marginal pointwise mutual informations [8]. This measure overcomes the interpretational issues of both $\mathrm{I}_{\mathrm{min}}$ and $\mathrm{I}_{\mathrm{BROJA}}$ and it can be applied to antichains of any size. Unfortunately, it does so at the expense of negativity, though one can argue that this is an accurate assessment of the information architecture.

With these measures, their approaches, and their limitations in mind, we now turn to defining our measure of unique information.

III. UNIQUE INFORMATION

We now propose a method to quantify partial information terms of the form $\mathrm{I}_\partial\left[i \to Y\right]$ —that is, the unique information. We begin by discussing the notion of *dependencies* and how to quantify their influence on information measures. We then adapt this to quantify how source-target dependencies influence the sources-target mutual information. Our measure quantifies unique information $\mathrm{I}_\partial\left[i \to Y\right]$ as the least amount that the $X_i Y$ dependency can influence $\mathrm{I}\left[X_{0:n} : Y\right]$.

A. Constraint Lattice

We begin by defining the *constraint lattice* $\mathcal{L}(\Sigma)$, a lattice of sets of subsets of variables. In this lattice no subset of variables implies another and each variable is represented at least once. Specifically, given a set of variables $\Sigma = \{X_0, X_1, \ldots\}$, a *constraint* γ is a nonempty subset of Σ. And, a *constraint set* σ is a set of constraints that form

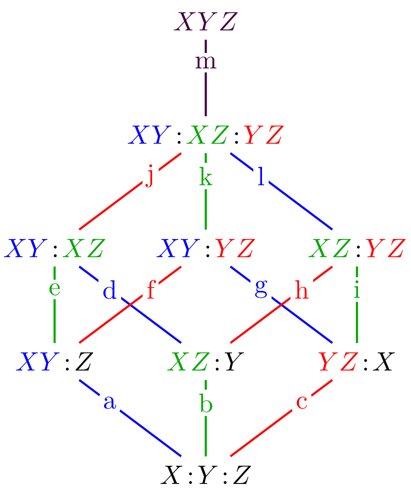


FIG. 2. Constraint lattice of three random variables X, Y, and Z. Blue edges (a, d, g, l) correspond to adding constraint XY, green (b, e, i, k) to XZ, and red (c, f, h, j) to YZ.

an antichain on Σ and whose union covers Σ; they are *antichain covers* [2]. Concretely, $\sigma \in \mathcal{P}^+(\mathcal{P}^+(\Sigma))$ such that, for all $\gamma_1, \gamma_2 \in \sigma$, $\gamma_1 \nsubseteq \gamma_2$ and $\bigcup \sigma = \Sigma$. The constraint sets are required to be covers since we are not concerned with each individual variable's distribution, rather we are concerned with how the variables are related. We refer to these variable sets as constraints since we will work with families of distributions for which marginal distributions over the variable sets are held fixed.

There is a natural partial order $\sigma_1 \preceq \sigma_2$ over constraint sets: for all $\gamma_1 \in \sigma_1$, there exists $\gamma_2 \in \sigma_2$ such that $\gamma_1 \subseteq \gamma_2$. The lattice $\mathcal{L}(\Sigma)$ induced by the partial order on $\Sigma = \{X, Y, Z\}$ is displayed in Fig. 2. The intuition going forward is that each node (antichain) in the lattice represents a set of constraints on marginal distributions and the constraints at one level imply those lower in the lattice.

B. Quantifying Dependencies

To quantify how each constraint set influences a distribution p, we associate a maximum entropy distribution with each constraint set σ in the lattice. Specifically, consider the set $\Delta_p(\sigma)$ of distributions that match marginals in σ with p:

$$\Delta_p(\sigma) = \{q : p(\gamma) = q(\gamma),~ \gamma \in \sigma\} ~. \tag{8}$$

To each constraint set σ we associate the distribution in $\Delta_p(\sigma)$ with maximal Shannon entropy:

$$p_\sigma = \arg\max \{\mathrm{H}[q] : q \in \Delta_p(\sigma)\} ~. \tag{9}$$

This distribution includes no additional biases beyond those constrained by σ [20]. When an information measure, such as the mutual information, is computed relative to this maximum entropy distribution, we will subscript it with the constraint: $\mathrm{I}_\sigma[XY : Z]$.

Given this lattice of maximum entropy distributions, we can then compute any multivariate information measure on those distributions and analyze how its value changes moving across the lattice. Moves here correspond to adding or subtracting dependencies. We call the lattice of information measures applied to the maximum entropy distributions the *dependency structure* of distribution p.

The dependency structure of a distribution is a flexible and robust method for analyzing how the structure of a distribution effects its information content. It allows each dependency to be studied in the context of other dependencies, leading to a vastly more nuanced view of the interactions among the variables. We believe it will form the basis for a wide variety of information-theoretic dependency analyses in the future.

C. Quantifying Unique Information

To measure the unique information that a source—say, X_0—has about the target Y, we use the dependency decomposition constructed from the mutual information between sources and the target. Consider further the lattice edges that correspond to the addition of a particular constraint:

$$E(\gamma) = \{(\sigma_1, \sigma_2) : (\sigma_1, \sigma_2) \in \mathcal{L}, \gamma \in \sigma_1, \gamma \notin \sigma_2\} ~. \tag{10}$$

For example, in Fig. 2's constraint lattice $E(XY)$ consists of the following edges: $(XY\!:\!Z, X\!:\!Y\!:\!Z)$, $(XY\!:\!XZ, XZ\!:\!Y)$, $(XY\!:\!YZ, YZ\!:\!X)$, and $(XY\!:\!XZ\!:\!YZ, XZ\!:\!YZ)$. These edges—labeled a, e, g, and l—are colored blue there. We denote a change in information measure along edge (σ_1, σ_2) by $\Delta^{\sigma_1}_{\sigma_2}$. For example, $\Delta^{\sigma_1}_{\sigma_2} \mathrm{I}[XY : Z] = \mathrm{I}_{\sigma_1}[XY : Z] - \mathrm{I}_{\sigma_2}[XY : Z]$.

Our measure $\mathrm{I}_{\text{dep}}[i \to Y]$ of unique information from variable X_i to the target Y is then defined using the lattice $\mathcal{L}(X_i, Y, X_{0:n \setminus i})$:

$$\mathrm{I}_{\text{dep}}[i \to Y] = \min_{(\sigma_1, \sigma_2) \in E(X_i Y)} \left\{\Delta^{\sigma_1}_{\sigma_2} \mathrm{I}[X_{0:n} : Y]\right\} ~. \tag{11}$$

That is, the information learned uniquely from X_i is the least change in sources-target mutual information among all the edges that involve the addition of the $X_i Y$ constraint. In the case of bivariate inputs, this measure of unique information results in a decomposition that satisfies **(S)**, **(SR)** (by construction), **(M)**, **(LP)**, and **(Id)**; see Appendix C for proofs.

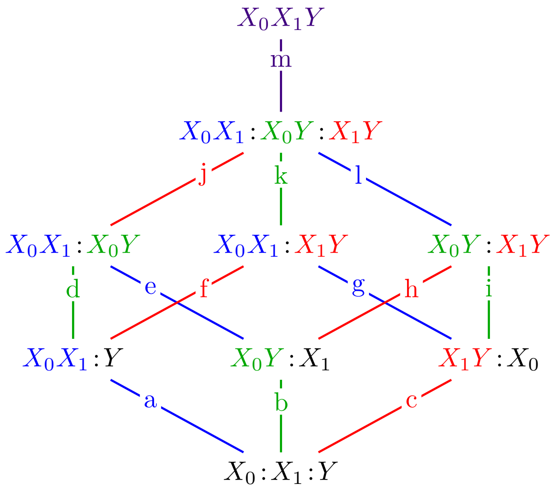


FIG. 3. Dependency structure for three variables X_0, X_1, and Y. Edges colored blue correspond to adding constraint $X_0\ X_1$; edges colored green to adding constraint $X_0\ Y$; and edges colored red to $X_1\ Y$. The unique information $\mathrm{I}_{\mathrm{dep}}[X_0 \to Y]$ is calculated by considering the least change in $\mathrm{I}_\sigma[X_0X_1 : Y]$ along the green edges. See Appendix B and Fig. 6 for identities among the edges important for $\mathrm{I}_{\mathrm{dep}}$.

Sum				Boom				Reduced Or(p)			
X_0	X_1	Y	Pr	X_0	X_1	Y	Pr	X_0	X_1	Y	Pr
0	0	0	1/4	0	0	1	1/6	0	0	0	1/2
0	1	1	1/4	0	0	2	1/6	0	0	1	p/4
1	0	1	1/4	0	2	0	1/6	0	1	1	$(1-p)$/4
1	1	2	1/4	1	2	1	1/6	1	0	1	$(1-p)$/4
				2	0	2	1/6	1	1	1	p/4
				2	1	2	1/6				

TABLE I. Three distributions of interest.

With a measure of unique information in hand, we now need only describe how to fill out the partial information lattice. In the bivariate sources case, this is straightforward: self-redundancy **(SR)**, the unique partial information values, and the Möbius inversion Eq. (2) complete the lattice. In the multivariate case, this is not generally possible, though in many relatively simple cases combining monotonicity **(M)**, self-redundancy **(SR)**, the unique values, and a few heuristics allow the lattice to be filled out. The heuristics include using the Möbius inversion on a subset of the lattice as a linear constraint. Several techniques such as this are implemented in the Python information theory package `dit` [21].

IV. EXAMPLES & COMPARISONS

We now demonstrate the behavior of our measure on a variety of source-target examples. In simple cases—Rdn, Syn, Copy [4]— $\mathrm{I}_{\mathrm{dep}}$ agrees with $\mathrm{I}_{\mathrm{proj}}$, $\mathrm{I}_{\mathrm{BROJA}}$, and $\mathrm{I}_{\mathrm{ccs}}$. There are, however, distributions where $\mathrm{I}_{\mathrm{dep}}$ differs from the rest.

Sum	$\mathrm{I}_{\mathrm{min}}$	$\mathrm{I}_{\mathrm{proj}}$	$\mathrm{I}_{\mathrm{BROJA}}$	$\mathrm{I}_{\mathrm{ccs}}$	$\mathrm{I}_{\mathrm{dep}}$
01	1	1	1	1/2	0.68872
0	0	0	0	1/2	0.31128
1	0	0	0	1/2	0.31128
0 · 1	1/2	1/2	1/2	0	0.18872

TABLE II. Partial information decomposition of the Sum distribution.

Boom	$\mathrm{I}_{\mathrm{min}}$	$\mathrm{I}_{\mathrm{proj}}$	$\mathrm{I}_{\mathrm{BROJA}}$	$\mathrm{I}_{\mathrm{ccs}}$	$\mathrm{I}_{\mathrm{dep}}$
01	0.29248	0.29248	0.12581	0.12581	0.08781
0	1/6	1/6	1/3	1/3	0.37133
1	1/6	1/6	1/3	1/3	0.37133
0 · 1	1/2	1/2	1/3	1/3	0.29533

TABLE III. Partial information decomposition of the Boom distribution.

Consider the Reduced Or(0) and Sum distributions [8] in Table I. For these $\mathrm{I}_{\mathrm{min}}$, $\mathrm{I}_{\mathrm{proj}}$, and $\mathrm{I}_{\mathrm{BROJA}}$ all compute no unique information. Reference [8] provides an argument based on game theory that the channels $X_0 \Rightarrow Y$ and $X_1 \Rightarrow Y$ being identical does not imply that unique information must vanish. (This is a special case of the Blackwell property **(BP)** [14].) Specifically, the argument goes, the optimization performed in computing $\mathrm{I}_{\mathrm{BROJA}}$ *artificially correlates* the sources. One can interpret this as a sign that redundancy is overestimated. In these instances, $\mathrm{I}_{\mathrm{dep}}$ qualitatively agrees with $\mathrm{I}_{\mathrm{ccs}}$, though they differ somewhat quantitatively. See Table II for the exact values.

Reference [5] proves that $\mathrm{I}_{\mathrm{proj}}$ and $\mathrm{I}_{\mathrm{BROJA}}$ are distinct measures. The only example produced, though, is the somewhat opaque Summed Dice distribution. Here, we offer Boom found in Table I as a more concrete example of this.[2] Interestingly, $\mathrm{I}_{\mathrm{min}}$ agrees with $\mathrm{I}_{\mathrm{proj}}$, while $\mathrm{I}_{\mathrm{ccs}}$ agrees with $\mathrm{I}_{\mathrm{BROJA}}$. $\mathrm{I}_{\mathrm{dep}}$, however, is distinct. In all cases, nonzero values are assigned to all four partial informations. Thus, it is not clear if any particular method is superior in this case. See Table III for the exact values.

Finally, consider the parametrized Reduced Or(p) distribution, seen in Table I. Figure 4 graphs this distribution's decomposition for all measures. $\mathrm{I}_{\mathrm{min}}$, $\mathrm{I}_{\mathrm{proj}}$, and $\mathrm{I}_{\mathrm{BROJA}}$ all produce the same decomposition as a function of p. $\mathrm{I}_{\mathrm{ccs}}$ and $\mathrm{I}_{\mathrm{dep}}$ differ from those three and each

[2] Although, it is not hard to find a simpler example: see the `dit` [21] documentation for another `http://docs.dit.io/en/latest/measures/pid.html#and-are-distinct`.

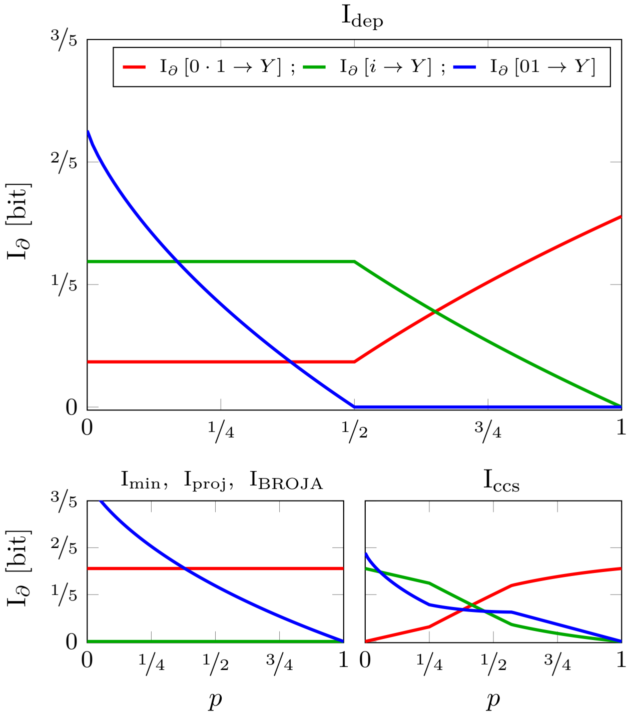


FIG. 4. Partial information decomposition of Reduced Or(p) as a function of p: The $\mathrm{I}_{\mathrm{dep}}$ decomposition shows an abrupt change in character at $p = 1/2$, corresponding to independent source-target channels switching from *underestimating* the target distribution to *overestimating*. Under $\mathrm{I}_{\mathrm{BROJA}}$ (and $\mathrm{I}_{\mathrm{min}}$ and $\mathrm{I}_{\mathrm{proj}}$), the redundant and unique components do not vary since the source-target marginals are invariant with p. The $\mathrm{I}_{\mathrm{ccs}}$ decomposition exhibits two p values of nonsmoothness, each corresponds to a change in sign of a coinformation component.

other. $\mathrm{I}_{\mathrm{BROJA}}$'s evaluation of redundant and unique information is invariant with p because the source-target marginals are fixed over p. We next demonstrate that $\mathrm{I}_{\mathrm{dep}}$'s decomposition is the most intuitive.

Generically, the source-target channels "overlap" independent of p, so there is some invariant component to the redundancy. Furthermore, the distribution $\Pr(Y) = \{0 : 1/2, 1 : 1/2\}$, while $\Pr(Y|X_0 = 0) = \Pr(Y|X_1 = 0) = \{0 : 2/3, 1 : 1/3\}$. Finally, $\Pr(Y|X_0X_1 = 00) = \{0 : 2/2{+}p, 1 : p/2{+}p\}$. If the two channels, $X_0 \Rightarrow Y$ and $X_1 \Rightarrow Y$, operated independently, then one would observe instead $\Pr(Y'|X_0X_1 = 00) = \{0 : 4/5, 1 : 1/5\}$. That is, each channel "pushes" $\Pr(Y = 0)$ $4/3$ of the way toward 1. This occurs exactly at $p = 1/2$. For $p \geq 1/2$, this independence assumption *overestimates* the probability of $Y = 0$. That is, there is additional redundancy between the two channels. For $p \leq 1/2$, the "pushes" from the two channels do not account for the true probability of $Y = 0$. That is, synergistic effects occur.

And

X_0	X_1	Y	Pr
0	0	0	1/4
0	1	0	1/4
1	0	0	1/4
1	1	1	1/4

TABLE IV. And distribution exemplifies both mechanistic redundancy and nonholistic synergy.

V. DISCUSSION

We next describe several strengths of our $\mathrm{I}_{\mathrm{dep}}$ measure in interpreting the behavior of the sources-target mapping channel. There are two aspects of the partial information decomposition that do not directly reflect properties of the joint distribution, but rather are resolved due to the selection of sources and target. The first involves redundancy, where two sources may be independent but redundantly influence the target. The second involves synergy, where there may be a lack of information at the triadic level of three-way interdependency, yet the sources synergistically influence the target. The dependency decomposition and $\mathrm{I}_{\mathrm{dep}}$ make these phenomena explicit.

A. Source versus Mechanistic Redundancy

An important question within the domain of partial information decomposition is that of *mechanistic redundancy*. This is redundant information that exceeds the mutual information of the sources. The And distribution seen in Table IV is a prototype for this phenomenon. Though the two sources X_0 and X_1 are independent, all methods of quantifying partial information ascribe nonzero redundancy to this distribution. Through the lens of $\mathrm{I}_{\mathrm{dep}}$, this occurs when the edge labeled l in Fig. 3 exceeds $b - i = c - h$. This means that the channels $X_0 \Rightarrow Y$ and $X_1 \Rightarrow Y$ are similar, so that when constraining just these two marginals the maximum entropy distribution artificially correlates the two sources. This artificial correlation must then be broken when constraining the sources marginal X_0X_1, leading to conditional dependence. (See Section V B for more on this implication.)

Mechanistic redundancy is closely tied to the concept of *target monotonicity* [14]:

$$\textbf{(TM)}\quad \mathrm{I}_\cap\,[X_0 \cdot X_1 \to Y] \geq \mathrm{I}_\cap\,[X_0 \cdot X_1 \to f(Y)] \quad .$$

(target monotonicity)

Said colloquially, taking a function of the target cannot increase redundancy. However, one of the following three properties must be false:

1. $\mathrm{I}_\cap\left[X_0 \cdot X_1 \rightarrow (X_0X_1)\right] = 0$,
2. Mechanistic redundancy, or
3. Target monotonicity .

$\mathrm{I}_{\mathrm{dep}}$ does not satisfy **(TM)**. Reference [14] demonstrated a general construction that maps a redundancy measure not satisfying **(TM)** to one that does, violating Item 1 in the process.

B. Holistic vs Non-Holistic Synergy

A notion somewhat complementary to mechanistic redundancy is nonholistic synergy. Holistic synergy is the difference between $\mathrm{H}_{\mathrm{X_0X_1:X_0Y:X_1Y}}\left[X_0X_1Y\right]$ and $\mathrm{H}\left[X_0X_1Y\right]$. It is information in the distribution that is only constrained by the full triadic probability distribution, also known as the third-order connected information [22]. This quantity appears as the edge labeled m in Fig. 3. Nonholistic synergy, on the other hand, is synergy that exists purely from the bivariate relationships within the distribution. This appears as $k - \min\{b, i, k\} = j - \min\{c, h, j\}$ in Fig. 3. This quantity has a natural interpretation: how much does the constraint X_0Y influence $\mathrm{I}\left[X_0X_1 : Y\right]$ in the context of the other dyadic relationships (X_0X_1, X_1Y), minus the unique information $\mathrm{I}_{\mathrm{dep}}\left[0 \rightarrow Y\right]$. The total PID synergy is then $\mathrm{I}_{\mathrm{dep}}\left[01 \rightarrow Y\right] = m + k - \min\{b, i, k\} = m + j - \min\{c, h, j\}$.

Here, again the AND distribution seen in Table IV exemplifies the phenomenon. The AND distribution is completely defined by the constraint $X_0X_1 : X_0Y : X_1Y$. This implies that the holistic synergy is zero for all information measures. In spite of this, all methods of quantifying partial information (correctly) assign nonzero synergy to this distribution. This is a consequence of coinformation being negative. This raises an interesting question: are there triadic (three-way) dependencies in the AND distribution? Notably, the distribution can be defined as the maximum entropy distribution satisfying certain pairwise marginals, yet it has negative coinformation and therefore nonzero synergy and exhibits conditional dependence.

VI. CONCLUSION

We developed a promising new method $\mathrm{I}_{\mathrm{dep}}$ of quantifying the partial information decomposition that circumvents many problems plaguing previous attempts. It satisfies axioms **(S)**, **(SR)**, **(M)**, **(LP)**, and **(Id)**; see Appendix C. It does not, however, satisfy **(BP)** and so, like $\mathrm{I}_{\mathrm{ccs}}$, it agrees with previous game-theoretic arguments raised in Ref. [8]. Unlike $\mathrm{I}_{\mathrm{ccs}}$, though, $\mathrm{I}_{\mathrm{dep}}$ satisfies **(LP)**. This makes it the only measure satisfying **(Id)** and **(LP)** that does not maintain that redundancy is fixed by $X_0Y : X_1Y$.

The $\mathrm{I}_{\mathrm{dep}}$ method does not overcome the negativity arising in the trivariate source explored in Refs. [9, 10, 13]. We agree with Ref. [13] that the likely solution is utilizing a different lattice. We further believe that the flexibility of our dependency structure could lead to methods of quantifying this hypothetical new lattice and to elucidating many other challenges in decomposing joint information.

ACKNOWLEDGMENTS

The authors thank Robin Ince for many helpful conversations. JPC thanks the Santa Fe Institute for its hospitality during visits. JPC is an SFI External Faculty member. This material is based upon work supported by, or in part by, the UC Davis Intel Parallel Computing Center, John Templeton Foundation grant 52095, Foundational Questions Institute grant FQXi-RFP-1609, the U.S. Army Research Laboratory and the U. S. Army Research Office under contracts W911NF-13-1-0390 and W911NF-13-1-0340.

[1] P. L. Williams and R. D. Beer. Nonnegative decomposition of multivariate information. *arXiv:1004.2515*. 1, 2, 3

[2] G. Birkhoff. *Lattice Theory*. American Mathematical Society, Providence, third edition, 1967. 1, 2, 4

[3] M. Harder, C. Salge, and D. Polani. Bivariate measure of redundant information. *Phys. Rev. E*, 87(1):012130, 2013. 1, 3

[4] V. Griffith and C. Koch. Quantifying synergistic mutual information. In *Guided Self-Organization: Inception*, pages 159–190. Springer, 2014. 3, 5

[5] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. 1, 3, 5

[6] D. Chicharro. Quantifying multivariate redundancy with maximum entropy decompositions of mutual information. *arXiv:1708.03845.*

[7] V. Griffith, E. K.P. Chong, R. G. James, C. J. Ellison, and J. P. Crutchfield. Intersection information based on common randomness. *Entropy*, 16(4):1985–2000, 2014. 1, 3

[8] R. A.A. Ince. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy*, 19(7):318, 2017. 1, 3, 5, 7

[9] N. Bertschinger, J. Rauh, E. Olbrich, and J. Jost. Shared information-new insights and problems in decomposing information in complex systems. In *Proceedings of the European Conference on Complex Systems 2012*, pages 251–269. Springer, 2013. 1, 3, 7

[10] J. Rauh, N. Bertschinger, E. Olbrich, and J. Jost. Reconsidering unique information: Towards a multivariate information decomposition. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 2232–2236. IEEE, 2014. 7

[11] D. Chicharro and S. Panzeri. Redundancy and synergy in dual decompositions of mutual information gain and information loss. *arXiv:1612.09522.* 1

[12] P. K. Banerjee and V. Griffith. Synergy, redundancy and common information. *arXiv:1509.03706.* 1

[13] J. Rauh. Secret sharing and shared information. *arXiv:1706.06998.* 1, 7

[14] J. Rauh, P. K. Banerjee, E. Olbrich, J. Jost, and N. Bertschinger. On extractable shared information. *arXiv:1701.07805.* 1, 5, 6, 7

[15] K. Krippendorff. Ross Ashby's information theory: a bit of history, some solutions to problems, and what we face today. *Intl J. General Systems*, 38(2):189–212, 2009. 1

[16] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, New York, second edition, 2006. 1

[17] D. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, Cambridge, United Kingdom, 2003.

[18] R. W. Yeung. *Information theory and network coding.* Springer, New York, 2008. 1

[19] A. J. Bell. The co-information lattice. In S. Makino S. Amari, A. Cichocki and N. Murata, editors, *Proc. Fifth Intl. Workshop on Independent Component Analysis and Blind Signal Separation*, volume ICA 2003, pages 921–926, New York, 2003. Springer. 3

[20] E. T. Jaynes. Where do we stand on maximum entropy? In E. T. Jaynes, editor, *Essays on Probability, Statistics, and Statistical Physics*, page 210. Reidel, London, 1983. 4

[21] R. G. James, C. J. Ellison, and J. P. Crutchfield. https://github.com/dit/dit: 1.0.0.dev8, September 2017. 5

[22] E. Schneidman, S. Still, M. J. Berry, W. Bialek, et al. Network information and connected correlations. *Phys. Rev. Lett.*, 91(23):238701, 2003. 7

Appendix A: Constrained Three-Variable Maximum Entropy Distributions

We derive properties of the constrained maximum entropy distributions that exist within the dependency structure of Section III B. These properties manifest themselves as restrictions on the variety of lattice edges that appear there. Consider a joint distribution $p(ABC)$ and maximum entropy distributions for some constraint set σ: $p_\sigma(ABC)$.

No pairwise marginal constraints

Consider distributions:

$$p_{A:B:C}(ABC) = p(A)p(B)p(C) \tag{A1}$$

With only single-variable marginal distributions constrained, the maximum entropy distribution is such that the variables are independent. Due to the marginal constraints, any increase in mutual information must correspond to a decrease in at least two conditional entropies. This decreases the total entropy and rules out the distribution as having maximum entropy. The information quantities follow from the independence of the variables.

One pairwise marginal constraint

Consider:

$$p_{AB:C}(ABC) = p(AB)p(C) \ . \tag{A2}$$

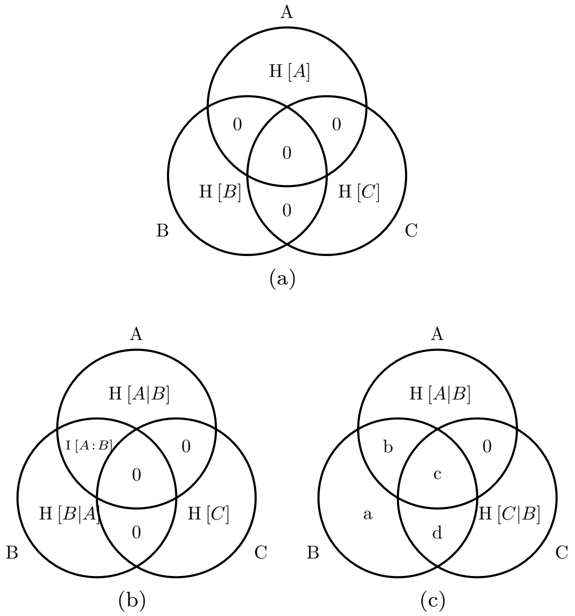
All information diagram atoms in the overlap of AB and C vanish. Any deviation from the information partitioning seen in Fig. 5b, which satisfies the constraint $AB : C$, must result in an overall decrease to the entropy. And so, the maximum entropy distribution is not $p_{AB:C}(ABC)$.

Two pairwise marginal constraints

Consider distributions obeying:

$$p_{AB:BC}(ABC) = p(A|B)p(B)p(C|B) \tag{A3}$$

Specifically, this forms a Markov chain $A - B - C$ and therefore $\mathrm{I}_{\mathrm{AB:BC}}[A:C|B] = 0$. With two two-variable marginal distributions constrained, there is no mutual information between the variables in only one constraint each, when conditioned on the variable that is in both constraints. In the expansion:



(a)

(b)

(c)

FIG. 5. Information diagrams corresponding to the maximum entropy distributions described in Eqs. (A1) to (A3). The four variables in subfigure (c) satisfy $a + b + c + d = \mathrm{H}[B]$, $b + c = \mathrm{I}[A:B]$, and $c + d = \mathrm{I}[B:C]$.

$$\mathrm{H}[ABC] = \mathrm{H}[B] + \mathrm{H}[C|AB] + \mathrm{H}[A|BC] + \mathrm{I}[A:C|B] \ ,$$

the first term is constrained and the remaining three terms have only one degree of freedom due to constraints on $p(C|B)$ and $p(A|B)$. Conditional mutual information is nonnegative, so the maximum possible entropy is that with zero conditional mutual information. Such a distribution is realized by the given Markov chain and it is, therefore, the maximum entropy distribution.

The information diagrams for each of these three distributions are given in Fig. 5. They are important to the following simplifications of the sources-target mutual information dependency structure.

Appendix B: Sources-Target Dependency Structure

Interpreting the set of antichain covers as possible distribution constraints, we defined a dependency lattice that introduces marginal dependencies into an otherwise independent, unbiased distribution. In this construction of a partial information decomposition, $\mathrm{I}_{\mathrm{dep}}$ is defined according to the node-node differences of sources-target mutual informations. These lattice edges are labeled in Fig. 3. Here, we show relationships arising among the edges summarized in Fig. 6.

The bottom of the dependency lattice constrains only the

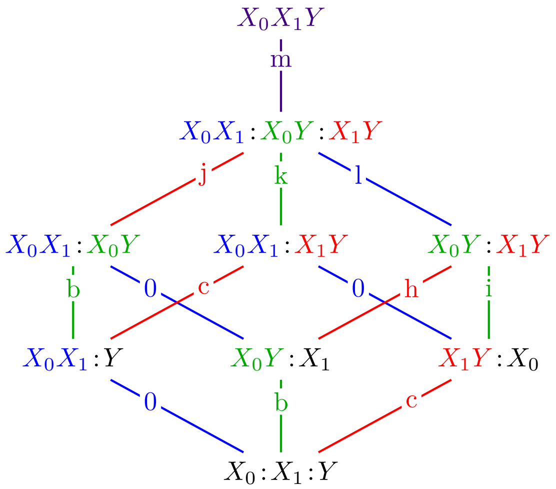


FIG. 6. The reduced form of the dependency lattice quantified by sources-target mutual information, with $b = \mathrm{I}[X_0 : Y]$ and $c = \mathrm{I}[X_1 : Y]$. All edges are guaranteed to be nonnegative except for l.

independent variables and has the maximum entropy distribution of Eq. (A1). Given appropriate permutations of pairs of nodes to be constrained, we use Eq. (A2) to see that $a = 0$, $b = \mathrm{I}[X_1 : Y]$ and $c = \mathrm{I}[X_1 : Y]$.

Similarly, appropriate permutations of Eqs. (A2) and (A3) give $e = g = 0$, as well as $d = b = \mathrm{I}[X_0 : Y]$ and $f = c = \mathrm{I}[X_1 : Y]$.

Considering nodes below edges h, i, j, and k, Eqs. (A2) and (A3) show that $\mathrm{I}_\sigma\left[X_i : Y | X_{\bar{i}}\right] = 0$ for an appropriate choice of i. Shifting across these edges by constraining $p(X_iY)$, then, any change in sources-target mutual information comes only from increases in this nonnegative conditional mutual information. Each of these four differences must therefore be nonnegative.

With all three pairwise marginal distributions constrained (and implicitly all three single variable marginal distributions constrained), the information diagram has only one free variable to be used to maximize entropy. In fact, maximizing the entropy is in this case is equivalent to minimizing the sources-target mutual information. It has been shown that further constraining the three-variable distribution must decrease the maximum entropy. This equivalently increases the sources-target mutual information and so $m \geq 0$.

This leaves only edge l unconstrained.

Of course, there are additional relationships between the edges; they are not otherwise wholly free. Of particular note, we will use the fact that the sum of edges along any path between two nodes must be equal. Here, we list all relationships we use in the next section:

$$a = e = g = 0 \tag{B1}$$

$$d = b = \mathrm{I}[X_0 : Y] \tag{B2}$$

$$f = c = \mathrm{I}[X_1 : Y] \tag{B3}$$

$$b + h = c + i \tag{B4}$$

$$b + j + m = c + k + m = \mathrm{I}[X_0X_1 : Y] \tag{B5}$$

Appendix C: Bivariate Partial Information $\mathrm{I}_{\mathbf{dep}}$ Decomposition

The following sections establish the properties of the bivariate partial information decomposition induced by $\mathrm{I}_{\mathrm{dep}}$.

Self-redundancy

Property **(SR)**:

$$\mathrm{I}_\cap[0 \rightarrow Y] = \mathrm{I}[X_0 : Y] \ .$$

We take this axiom constructively, but the redundancy values for unique information are consistent with the simple one-source lattice. In a sense, this claims that $\mathrm{I}_\cap[X_0 \rightarrow Y]$ is invariant to the addition or removal of other input variables. This axiom is required for filling out the two-source redundancy lattice.

Monotonicity

Property **(M)**: For all α, β:

$$\alpha \preceq \beta \implies \mathrm{I}_\cap[\alpha \rightarrow Y] \leq \mathrm{I}_\cap[\beta \rightarrow Y] \ .$$

This follows directly from **(SR)** and the fact that $\mathrm{I}_{\mathrm{dep}}[0 \rightarrow Y] = \min\{b, i, k\}$, where $b = \mathrm{I}[X_0 : Y]$. That is, the source-target mutual information is one element of the set of which the unique information is the minimum.

Nonnegativity

Property **(LP)**: For all σ,

$$\mathrm{I}_\partial[\sigma \rightarrow Y] \geq 0 \ .$$

Begin by considering $\mathrm{I}_\partial[0 \rightarrow Y] = \min(b, i, k) \geq 0$. All arguments of this minimum have been shown to be nonnegative.

Using the self-redundancy axiom to define $\mathrm{I}_\cap\left[0 \to Y\right]$, we have:

$$\begin{aligned} \mathrm{I}_\partial\left[0 \cdot 1 \to Y\right] &= \mathrm{I}\left[X_0 : Y\right] - \mathrm{I}_\partial\left[0 \to Y\right] \\ &= b - \min\left(b, i, k\right) \\ &\geq 0 \ . \end{aligned}$$

To determine the remaining two partial information atoms, we must consider the ordering of $\{b, i, k\}$. We designate three overlapping cases: one for each of the possible minimum elements. Reductions are done by using the results of Appendix B. We repeatedly use the redundancy lattice inversions: $\mathrm{I}_\partial\left[1 \to Y\right] = \mathrm{I}_\cap\left[1 \to Y\right] - \mathrm{I}_\partial\left[0 \cdot 1 \to Y\right]$ and $\mathrm{I}_\partial\left[01 \to Y\right] = \mathrm{I}_\cap\left[01 \to Y\right] - \mathrm{I}_\partial\left[0 \cdot 1 \to Y\right] - \mathrm{I}_\partial\left[0 \to Y\right] - \mathrm{I}_\partial\left[1 \to Y\right]$.

CASE 1: $b \leq i,\ b \leq k$.

$$\begin{aligned} \mathrm{I}_\partial\left[1 \to Y\right] &= c - 0 \\ &\geq 0 \ , \\ \mathrm{I}_\partial\left[01 \to Y\right] &= (c + k + m) - 0 - b - c \\ &= m + k - b \\ &\geq m \\ &\geq 0 \ . \end{aligned}$$

CASE 2: $i \leq b,\ i \leq k$.

$$\begin{aligned} \mathrm{I}_\partial\left[1 \to Y\right] &= c - (b - i) = h \\ &\geq 0 \ , \\ \mathrm{I}_\partial\left[01 \to Y\right] &= (c + k + m) - (c - h) - i - h \\ &= m + k - i \\ &\geq m \\ &\geq 0 \ . \end{aligned}$$

CASE 3: $k \leq b,\ k \leq i$.

$$\begin{aligned} \mathrm{I}_\partial\left[1 \to Y\right] &= c - (b - k) \\ &= j \\ &\geq 0 \ , \\ \mathrm{I}_\partial\left[01 \to Y\right] &= (b + j + m) - (b - k) - k - j \\ &= m \\ &\geq 0 \ . \end{aligned}$$

In each case, the second unique information is found to be equivalent to another nonnegative edge in the dependency lattice. And, the synergy is found to be bounded from below by the "holistic" synergy m.

Symmetry

Property **(S)**: Under source reorderings, the following is invariant:

$$\mathrm{I}_\partial\left[0 \to Y\right] \ .$$

The dependency lattice is symmetric by design. Relabeling the random variables is equivalent to an isomorphic relabeling of the lattice. Therefore, we consider the effect of completing the partial information decomposition by either $\mathrm{I}_{\mathrm{dep}}\left[0 \to Y\right]$ or $\mathrm{I}_{\mathrm{dep}}\left[1 \to Y\right]$.

Computing $\mathrm{I}_{\mathrm{dep}}\left[0 \to Y\right] = \min\left(b, i, k\right)$ gives $\mathrm{I}_\partial\left[1 \to Y\right] = \min\left(c, h, j\right)$, although we never explicitly do the second minimization. This requires simple algebra from the various multiple-paths constraints given in Appendix B. In each of the cases in Appendix C, $\mathrm{I}_\partial\left[1 \to Y\right]$ was found to be one of $\{c, h, j\}$. Straightforward algebra shows that it is necessarily the minimum of them in each of the particular cases.

Identity

Property **(Id)**:

$$\mathrm{I}_\partial\left[0 \cdot 1 \to X_0 X_1\right] = \mathrm{I}\left[X_0 : X_1\right] \ .$$

Consider sources X_0, X_1 and output $Y = X_0 X_1$, the concatenation of inputs. The mutual information of either source with the target is simply the entropy of that source. That is, $b = \mathrm{H}\left[X_0\right]$. Using appropriate permutations of Eqs. (A2) and (A3) ($A = X_0$, $B = Y$, $C = X_1$), we find that $i = \mathrm{H}\left[X_0 | X_1\right]$. Now, starting with Eq. (A3) ($A = X_0$, $B = X_1$, $C = Y$), we see that additionally constraining $p(X_0 Y)$ fully constrains the distribution to its original form, with a sources-target mutual information of $\mathrm{H}\left[X_0 X_1\right]$. That is, $k = \mathrm{H}\left[X_0 | X_1\right]$. The minimum of these three quantities gives $\mathrm{I}_{\mathrm{dep}}\left[0 \to Y\right] = \mathrm{H}\left[X_0 | X_1\right]$ and therefore verifies the identity axiom.