Santa Fe Institute Working Paper 17-09-034
arXiv:1709.06653

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: October 27, 2018)

The partial information decomposition (PID) is perhaps the leading proposal for resolving information shared between a set of sources and a target into redundant, synergistic, and unique constituents. Unfortunately, the PID 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. We first develop a broadly applicable method—the dependency decomposition—that delineates how statistical dependencies influence the structure of a joint distribution. The dependency decomposition then allows us to define a measure of the information about a target that can be uniquely attributed to a particular source as the least amount which the source-target statistical dependency can influence the information shared between the sources and the target. The result is the first measure that satisfies the core axioms of the PID framework while not satisfying the Blackwell relation, which depends on a particular interpretation of how the variables are related. This makes a key step forward to a practical PID.

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 [1], neural coding [2], highly-correlated electron systems, spin lattices [3], financial markets [4], network design [5], and other complex systems whose large-scale organization is either not known a priori or emerges spontaneously. Information theory's originator Claude Shannon [6] was open to the possible benefits of such applications; he was also wary [7]. In an early attempt to lay common foundations for multicomponent, multivariate information Shannon [8] appealed to Garrett Birkhoff's lattice theory [9]. Many of the questions raised are still open today [10, 11].

Along these lines, but rather more recent, one particularly promising framework for accomplishing such a decomposition is the *partial information decomposition* (PID) [10]. 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 case of two sources: redundant, unique, and synergistic informations. This task relies on two separate aspects of the framework: first, the overlapping source subsets into which the information should be decomposed and, second, the method of quantifying those informational components.

Unfortunately, despite a great deal of effort [12–21], the current consensus is (i) that the lattice needs to be modified [18–20, 22, 23] and (ii) that extant methods of quantifying informational components [10, 12–14, 16, 17] are not satisfactory in full multivariate generality due to either only quantifying unique informations, being applicable only to two-source distributions, or lacking nonnegativity. Thus, the promise of a full informational analysis of the organization of complex systems remains unrealized after more than a half century.

The following addresses the second aspect—quantifying the components. Inspired by early cybernetics—specifically, Krippendorff's lattice of system models (reviewed in Ref. [24])—we develop a general technique for decomposing arbitrary multivariate information measures according to how they are influenced by statistical dependencies.[1] We then use this decomposition to quantify the information that one variable uniquely has about another. Reference [14]'s I_{BROJA} measure also directly quantifies unique information. However, depending upon one's intuitions [26] it can be seen to inflate redundancy [17]. Both our measure as well as Ref. [17]'s

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

[1] Since the development of this manuscript, it has come to the authors' attention that this structure had been independently developed within the field of system science [25].

$\mathrm{I}_{\mathrm{ccs}}$ take into account the joint statistics of the sources, 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 nonnegative, respects the source statistics, and satisfies the core axioms of the PID framework.

Our development proceeds as follows. Section II reviews the PID 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 [27–29].

II. BACKGROUND

Consider a set of *sources* $X_0, X_1, \ldots, X_{n-1} = X_{0:n}$ and a *target* Y.[2] The amount of information the sources carry about the target is quantified by their mutual information:

$$\begin{aligned}\mathrm{I}\left[X_{0:n} : Y\right] &= \mathrm{I}\left[X_0, X_1, \ldots, X_{n-1} : Y\right] \\ &= \sum p(X_{0:n}, Y) \log_2 \frac{p(X_{0:n}, Y)}{p(X_{0:n})p(Y)} \ .\end{aligned}$$

The PID then assigns *shared information* to sets of source groupings such that no (inner) set is subsumed by another [10]. 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 lattice [9], where one antichain α is less than another β if each element in β subsumes some element of α:

$$\alpha \preceq \beta \iff \forall s_1 \in \beta, \exists s_2 \in \alpha, s_2 \subseteq s_1 \ .$$

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 $\{\{X_0\}\{X_1, X_2\}\}$.

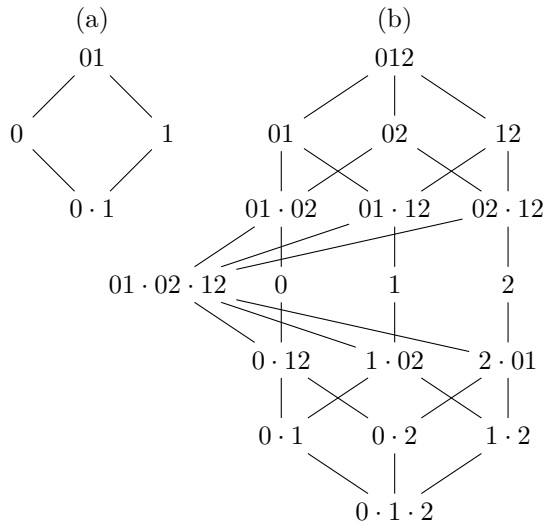

FIG. 1. Lattice of antichains for (a) two (X_0 and X_1) and (b) three sources (X_0, X_1, and X_2): An antichain is represented using a dot to separate sets and sets by concatenated indices; *e.g.*, $\{\{X_0\}\{X_1, X_2\}\}$ is represented $0 \cdot 12$.

B. Shared Informations

Given the antichain lattice, one then assigns a quantity of shared or redundant information to each antichain. 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 $\mathrm{I}_\cap\left[\alpha \to Y\right]$. Reference [10] put forth several axioms that such a measure should follow:

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

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

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

With a lattice of shared informations in hand, the *partial information* $\mathrm{I}_\partial\left[\alpha \to Y\right]$ is defined as the Möbius inversion [9] of the shared information:

$$\mathrm{I}_\cap\left[\alpha \to Y\right] = \sum_{\beta \preceq \alpha} \mathrm{I}_\partial\left[\beta \to Y\right] \ . \tag{1}$$

We further require that the following axiom hold:

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

[2] We subscript the joint variable with a Python-like array-slice notation, which matches Dijkstra's argument [30].

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

C. The Bivariate Case

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

$$\begin{aligned} \mathrm{I}\left[X_0X_1 : Y\right] = \mathrm{I}_\partial\left[0 \cdot 1 \to Y\right] + \mathrm{I}_\partial\left[0 \to Y\right] \\ + \mathrm{I}_\partial\left[1 \to Y\right] + \mathrm{I}_\partial\left[01 \to Y\right] \ , \end{aligned} \tag{2}$$

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

$$\mathrm{I}\left[X_0 : Y\right] = \mathrm{I}_\partial\left[0 \cdot 1 \to Y\right] + \mathrm{I}_\partial\left[0 \to Y\right] \tag{3}$$

$$\mathrm{I}\left[X_1 : Y\right] = \mathrm{I}_\partial\left[0 \cdot 1 \to Y\right] + \mathrm{I}_\partial\left[1 \to Y\right] \ . \tag{4}$$

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

$$\begin{aligned} \mathrm{I}\left[X_0 : Y | X_1\right] &= \mathrm{I}\left[X_0X_1 : Y\right] - \mathrm{I}\left[X_1 : Y\right] \\ &= \mathrm{I}_\partial\left[0 \to Y\right] + \mathrm{I}_\partial\left[01 \to Y\right] \ . \end{aligned}$$

Furthermore, the co-information [31] can be expressed as:

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

This illustrates one of the PID's strengths. 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. (2) to (4). This leaves one degree of freedom. Generally, though not always [14], this is taken as $\mathrm{I}_\partial\left[0 \cdot 1 \to Y\right]$. Therefore, specifying any component of the partial information lattice determines the entire decomposition. In the multivariate case, however, no single $\mathrm{I}_\partial\left[\alpha \to Y\right]$ (redundancy, unique, synergistic, or otherwise), when combined with relations to standard information-theoretic quantities, will determine the remainder of the values. For this reason, one generally relies upon quantifying the $\mathrm{I}_\cap$ values to complete the decomposition via the Möbius inversion.

Finally, in the bivariate case one further axiom has been suggested [12], though not put forth in original PID:

(Id) $\mathrm{I}_\cap\left[0 \cdot 1 \to X_0X_1\right] = \mathrm{I}\left[X_0 : X_1\right]$ *(identity)*

This axiom ensures that simply concatenating independent inputs does not result in redundant information. The identity axiom, though intuitive in the case of two inputs, suffers from several issues. Primarily, with three or more sources it is known to be inconsistent with local positivity **(LP)**. Furthermore, it is not clear how to extend this axiom to the multivariate case or even if it should be extended. In short, though many proposed methods of quantifying the PID satisfy the identity axiom, it is certainly not universally accepted.

D. Extant Methods

Several methods can be easily set aside— $\mathrm{I}_{\mathrm{mmi}}$ [18], $\mathrm{I}_\wedge$ [16], and $\mathrm{I}_\downarrow$ [13, 18]—as suffering from significant drawbacks. $\mathrm{I}_{\mathrm{mmi}}$ necessarily assigns a zero value to at least one of the unique informations, doing so by dictating that whichever source shares the least amount of information with the target, it does so entirely redundantly with the other sources. $\mathrm{I}_\wedge$ is based on the Gács-Körner common information. And, so it is insensitive to any sort of statistical correlation that is not a common random variable. $\mathrm{I}_\downarrow$ quantifies unique information from each source directly using an upper bound on the secret key agreement rate [32], but in a way that leads to inconsistent redundancy and synergy values.

Now, we can turn to describe the four primary existing methods for quantifying the PID. $\mathrm{I}_{\mathrm{min}}$, the first measure proposed [10], quantifies the average least information the individual sources have about each target value. It has been criticized [12, 13] for its behavior in certain situations. For example, when the target simply concatenates two independent bit-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 [12]. Due to its foundation relying on the

Kullback-Leibler divergence, however, it does not naturally generalize to measuring the shared information in antichains of size three or greater.

$\mathrm{I}_{\mathrm{BROJA}}$ attempts to quantify unique information [14], as does our approach. 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 X_i.) Depending on one's perspective on the roles of source and target variables, however, $\mathrm{I}_{\mathrm{BROJA}}$ can be seen to artificially correlate the sources and thereby overestimate redundancy [17, 26]. 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 [17]. This measure seems to avoid the issue used to criticize $\mathrm{I}_{\mathrm{min}}$, that of capturing the "same quantity" rather than the "same information". Further, it respects the source statistics unlike $\mathrm{I}_{\mathrm{BROJA}}$. Notably, it can be applied to antichains of any size. Unfortunately, it does so incurring the expense of negativity, though one can argue that this is an accurate assessment of 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 based on 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 defines unique information $\mathrm{I}_\partial\left[i \to Y\right]$ as the least amount that the $X_i Y$ dependency can influence the shared information $\mathrm{I}\left[X_{0:n} : Y\right]$.

A. Constraint Lattice

We begin by defining the *constraint lattice* $\mathcal{L}(\Sigma)$, a lattice of sets of subsets of variables. These subsets of variables are antichains, as in the partial information lattice, but are further constrained and endowed with a different ordering. 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 an antichain on Σ and whose union covers Σ; they are *antichain covers* [9]. Concretely, $\sigma \in \mathcal{P}^+(\mathcal{P}^+(\Sigma))$ such that, for all $\gamma_1, \gamma_2 \in \sigma$, $\gamma_1 \not\subseteq \gamma_2$ 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 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:

$$\sigma_1 \preceq \sigma_2 \iff \forall \gamma_1 \in \sigma_1, \exists \gamma_2 \in \sigma_2, \gamma_1 \subseteq \gamma_2 \ .$$

Note that this relation is somewhat dual to that in the PID and, furthermore, that the set of antichain covers is a subset of all antichains. 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:

$$\Delta_p(\sigma) = \{q : p(\gamma) = q(\gamma),\ \gamma \in \sigma\} \ . \tag{5}$$

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

$$p_\sigma = \arg\max \{\mathrm{H}\left[q\right] : q \in \Delta_p(\sigma)\} \ . \tag{6}$$

This distribution includes no additional statistical structure beyond that constrained by σ [33].

When an information measure, such as the mutual information, is computed relative to the maximum entropy distribution p_σ, we subscript it with the constraint σ. For example, the mutual information between the joint variable XY and the variable Z relative the the maximum entropy distribution satisfying the constraint $XY : YZ$ is denoted $\mathrm{I}_{\mathrm{XY:YZ}}\left[XY : Z\right]$.

Given this lattice of maximum entropy distributions, we can then compute any multivariate information measure

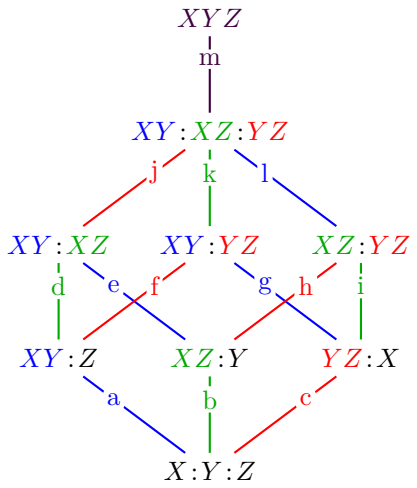

FIG. 2. Constraint lattice of three random variables X, Y, and Z. Blue edges (a, e, g, l) correspond to adding constraint XY, green (b, d, i, k) to adding XZ, and red (c, f, h, j) to adding YZ.

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 broadly applicable and robust method for analyzing how the structure of a distribution affects its information content. It is effectively a partial order on a multiverse associated with p: Consider every possible alternative universe in which select statistical dependencies are removed from p. It allows each dependency to be studied in the context of other dependencies, leading to a vastly more nuanced view of the interactions among p's variables. We believe it will form the basis for a wide variety of information-theoretic dependency analyses in the future.

We note that this dependency structure was independently announced [34] after a preprint (arxiv.org:1609.01233) of the present work appeared. There, dependency structure minimizes the Kullback-Leibler divergence, which is known to be equivalent to our maximum entropy approach [35]. A decomposition of total correlation $\sum_j \mathrm{H}[X_i] - \mathrm{H}[X_1, \ldots, X_n]$ was studied that, in the case of three variables, amounts to decomposing each conditional mutual information into two components: e.g., $\mathrm{I}[X : Y|Z] = \mathrm{D}_{\mathrm{KL}}[XY : XZ : YZ \;||\; XZ : YZ] + \mathrm{D}_{\mathrm{KL}}[XYZ \;||\; XY : XZ : YZ]$, where the latter is the third-order connected information [36] and $\mathrm{D}_{\mathrm{KL}}[P \;||\; Q]$ is the Kullback-Liebler divergence [27]. In contrast to a lattice decomposition of total correlations, the primary contribution here applies any desired information measure to each node in the dependency structure. This leads to a vast array of possible analyses.

As an example, consider the problem of determining "causal pathways" in a network [37, 38].[3] Given two paths between two network nodes, say $A \to B \to C$ and $A \to B' \to C$, one would like to determine through which pathway A's behavior most strongly influences C. This pathway is termed the causal pathway. Naïvely, one might assume that the pathway whose links are strongest is the more influential pathway or that it is the pathway maximizing a multivariate information measure. Consider, however, the case where A strongly influences one aspect of B while another, independent aspect of B strongly influences C. Here, A would have no influence whatsoever upon C in spite of the strong individual links. Our dependency structure, in contrast, would easily detect this via $\mathrm{I}_{\mathrm{AB:BC}}[A : C] = 0$, which quantifies exactly how much A and C are correlated through the pathway $A \to B \to C$. In this fashion, determining causal pathways in networks becomes straightforward: For each potential pathway, consider the constraint consisting of all its links and evaluate the influence measure of choice (time-delayed mutual information, transfer entropy, or similar) between the beginning and end of the pathway. The value indicates the pathway's strength as quantified by the chosen measure (and whose interpretation is dependent on the chosen measure). While this example does not use the full dependency structure, it does demonstrate the usefulness of considering information measures in the context of only specific dependencies. Furthermore, there exist other information-based methods of determining causal pathways [37], however this provides a novel and independent method of doing so.

Another application is the determination of drug interactions. Given a dataset of responses to a variety of drugs, one would like to determine which subsets of drugs interact with one another. One method of doing so would be to construct the dependency structure, quantifying each node with the entropy. Then, the lowest node in the lattice whose entropy is "close enough" (as determined by context) to that of the true distribution contains the minimal set of constraints that give rise to the full structure of the true distribution. That minimal set of constraints determines the subset of variables that are necessarily interacting. Note that this is an application of reconstructability analysis [25] and does not use the flexibility of employing a variety of information measures on the lattice.

[3] The term "causal" here is unfortunate, due to the great deal of debate on determining causality without intervention [39].

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:

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

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, d, 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, $\Delta^{\sigma_1}_{\sigma_2} \mathrm{I}\left[XY : Z\right] = \mathrm{I}_{\sigma_1}\left[XY : Z\right] - \mathrm{I}_{\sigma_2}\left[XY : Z\right]$.

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

$$\begin{aligned}\mathrm{I}_{\mathrm{dep}}\left[i \to Y\right] &= \min_{(\sigma_1,\sigma_2)\in E(X_iY)} \left\{\Delta^{\sigma_1}_{\sigma_2} \mathrm{I}\left[X_{0:n} : Y\right]\right\} \\ &= \min \begin{cases} \mathrm{I}_{\mathrm{X_i Y : X_{0:n\setminus i} Y}}\left[X_i : Y | X_{0:n\setminus i}\right] \\ \mathrm{I}_{\mathrm{X_i X_{0:n\setminus i} : X_i Y : X_{0:n\setminus i} Y}}\left[X_i : Y | X_{0:n\setminus i}\right] \end{cases} .\end{aligned}$$

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_iY constraint. Due to information-theoretic constraints (see Appendix A), the edge difference achieving the minimum must be one of either $\mathrm{I}_{\mathrm{X_i Y : X_{0:n\setminus i} Y}}\left[X_i : Y | X_{0:n\setminus i}\right]$ or $\mathrm{I}_{\mathrm{X_i X_{0:n\setminus i} : X_i Y : X_{0:n\setminus i} Y}}\left[X_i : Y | X_{0:n\setminus i}\right]$. This latter quantity arises in directed reconstructability analysis [25], where it has an interpretation similar to unique information. It would not, however, result in a PID that satisfied **(LP)**; though, when combined as above with the first quantity, local positivity is preserved. 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)**, as shown in Appendix C. In the case of multivariate inputs, satisfying **(Id)** implies that **(LP)** is not satisfied. Further, it is not clear whether $\mathrm{I}_{\mathrm{dep}}$ satisfies **(M)**.

With a measure of unique information in hand, we now need only describe how to determine 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. (1) complete the lattice. In the multivariate case, completion is not generally possible. That said, in many relatively simple cases combining monotonicity **(M)**, self-redundancy **(SR)**, the unique values, and a few heuristics allow the lattice to be determined. Though, due to $\mathrm{I}_{\mathrm{dep}}$ satisfying the identity axiom such values may violate local positivity. 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` [40].

SUM

X_0	X_1	Y	Pr
0	0	0	1/4
0	1	1	1/4
1	0	1	1/4
1	1	2	1/4

PNT. UNQ.

X_0	X_1	Y	Pr
0	1	1	1/4
1	0	1	1/4
0	2	2	1/4
2	0	2	1/4

BOOM

X_0	X_1	Y	Pr
0	0	1	1/6
0	0	2	1/6
0	2	0	1/6
1	2	1	1/6
2	0	2	1/6
2	1	2	1/6

REDUCED OR(p)

X_0	X_1	Y	Pr
0	0	0	1/2
0	0	1	p/4
0	1	1	$(1-p)$/4
1	0	1	$(1-p)$/4
1	1	1	p/4

FIG. 3. Four distributions of interest: SUM is constructed with X_0, X_1 as independent binary variables while Y is their sum. PNT. UNQ. is from Ref. [41]. BOOM was found through a numeric search for distributions satisfying certain properties; namely, that $\mathrm{I}_{\mathrm{proj}}$ and $\mathrm{I}_{\mathrm{BROJA}}$ differ. REDUCED OR is adapted from Ref. [17].

IV. EXAMPLES & COMPARISONS

We now demonstrate the behavior of our measure $\mathrm{I}_{\mathrm{dep}}$ on a variety of source-target examples. In simple cases—RDN, SYN, COPY [13]— $\mathrm{I}_{\mathrm{dep}}$ 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. We concentrate on those.

Consider the REDUCED OR(0) and SUM distributions [17] in Fig. 3. For these $\mathrm{I}_{\mathrm{min}}$, $\mathrm{I}_{\mathrm{proj}}$, and $\mathrm{I}_{\mathrm{BROJA}}$ all compute no unique information. Reference [17] provides an argument based on game theory that the channels $X_0 \Rightarrow Y$ and $X_1 \Rightarrow Y$ being identical (a special case of the Blackwell property **(BP)** [23]) does not imply that unique information must vanish. Specifically, the argument goes, the optimization performed in computing $\mathrm{I}_{\mathrm{BROJA}}$ *artificially correlates* the sources, though this interpretation is dependent upon the perspective one takes when considering the PID [26]. One can interpret this as a sign that redundancy is being overestimated.

	∂	$\mathrm{I}_{\min}$	$\mathrm{I}_{\mathrm{proj}}$	$\mathrm{I}_{\mathrm{BROJA}}$	$\mathrm{I}_{\mathrm{ccs}}$	$\mathrm{I}_{\mathrm{dep}}$
Sum	01	1	1	1	$1/2$	0.68872
	0	0	0	0	$1/2$	0.31128
	1	0	0	0	$1/2$	0.31128
	$0 \cdot 1$	$1/2$	$1/2$	$1/2$	0	0.18872
Pnt. Unq.	01	$1/2$	$1/2$	$1/2$	0	$1/4$
	0	0	0	0	$1/2$	$1/4$
	1	0	0	0	$1/2$	$1/4$
	$0 \cdot 1$	$1/2$	$1/2$	$1/2$	0	$1/4$
Boom	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
	$0 \cdot 1$	$1/2$	$1/2$	$1/3$	$1/3$	0.29533

TABLE I. Partial information decomposition of the Sum, Pnt. Unq., and Boom distributions.

In these instances, $\mathrm{I}_{\mathrm{dep}}$ qualitatively agrees with $\mathrm{I}_{\mathrm{ccs}}$, though they differ somewhat quantitatively. See Table I for the exact values.

Reference [14] 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 Fig. 3 as a more concrete example to draw out such differences.[4] Table I gives the measures' decomposition values. Interestingly, $\mathrm{I}_{\min}$ agrees with $\mathrm{I}_{\mathrm{proj}}$, while $\mathrm{I}_{\mathrm{ccs}}$ agrees with $\mathrm{I}_{\mathrm{BROJA}}$. $\mathrm{I}_{\mathrm{dep}}$, however, is distinct. All measures assign nonzero values to all four partial informations. Thus, it is not clear if any particular method is superior in this case.

Finally, consider the parametrized Reduced Or(p) distribution, given in Fig. 3. Figure 4 graphs this distribution's decomposition for all measures. $\mathrm{I}_{\min}$, $\mathrm{I}_{\mathrm{proj}}$, 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 other. $\mathrm{I}_{\min}$'s, $\mathrm{I}_{\mathrm{proj}}$'s, and $\mathrm{I}_{\mathrm{BROJA}}$'s evaluation of redundant and unique information is invariant with p. And, in the cases of $\mathrm{I}_{\mathrm{proj}}$ and $\mathrm{I}_{\mathrm{BROJA}}$ this is due to the source-target marginals being invariant with respect to p [14]. We next argue that $\mathrm{I}_{\mathrm{dep}}$'s decomposition—and specifically the "kink" observed—is both intuitive and reasonable.

Generically, the source-target channels "overlap" independent of p, so there is some invariant component to the redundancy. Furthermore, consider the form of some conditional distributions when the sources are '0': the distribution $\Pr(Y) = \{0 : 1/2, 1 : 1/2\}$, while

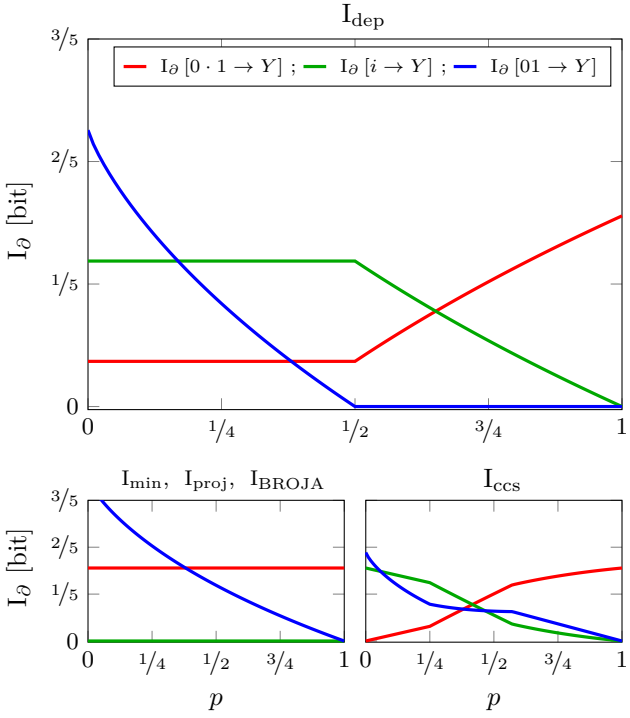

FIG. 4. Partial information decomposition of Reduced Or(p) as a function of p: The $\mathrm{I}_{\mathrm{dep}}$ decomposition shows an abrupt change in character at $p = 1/2$, corresponding to independent source-target channels switching from *underestimating* the target distribution to *overestimating*. Under $\mathrm{I}_{\mathrm{BROJA}}$ (and $\mathrm{I}_{\min}$ 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.

$\Pr(Y|X_0 = 0) = \Pr(Y|X_1 = 0) = \{0 : 2/3, 1 : 1/3\}$. Finally, $\Pr(Y|X_0X_1 = 00) = \{0 : 2/2+p, 1 : p/2+p\}$. The source-target channels are independent of p, but the joint sources-target channel depends upon it.

Consider the case of the two channels, $X_0 \Rightarrow Y$ and $X_1 \Rightarrow Y$, operating and independently influencing the value of Y. Observation that $X_0 = 0$ takes $\Pr(Y = 0) = 1/2$ to $\Pr(Y = 0) = 2/3$—a factor of $4/3$ larger, and similarly for $X_1 = 0$. Together, one would then observe that $\Pr(Y'|X_0X_1 = 00) = \{0 : 4/5, 1 : 1/5\}$. That is, each channel "pushes" $\Pr(Y = 0)$ $4/3$ of the way from '0' toward '1'. This independent "pushing" 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. $\mathrm{I}_{\mathrm{dep}}$ cleanly reveals both of these features, while $\mathrm{I}_{\min}$, $\mathrm{I}_{\mathrm{proj}}$, and $\mathrm{I}_{\mathrm{BROJA}}$ miss them and $\mathrm{I}_{\mathrm{ccs}}$'s multiple kinks makes it appear oversensitive.

[4] Although, it is not hard to find a simpler example. See the `dit` [40] documentation for another: `http://docs.dit.io/en/latest/measures/pid.html#and-are-distinct`.

V. DISCUSSION

We next describe several strengths of our $\mathrm{I}_{\mathrm{dep}}$ measure when interpreting the behavior of the sources-target mapping channel. The PID applied to a joint distribution naturally depends on selecting which variables are considered sources and which target. In some cases—the RDN and SYN distributions of Ref. [13]—the values of redundancy and synergy are independent of these choices and, in a sense, can be seen as a property of the joint distribution itself. In other cases—the AND distribution of Ref. [13]—the values of redundancy and synergy are not readily apparent in the joint distribution. The concept of mechanistic redundancy—the existence of redundancy in spite of independent sources—is a manifestation of this. What we term nonholistic synergy—synergy in the PID that does not arise from necessarily three-way interactions (that is, the third-order connected information [36]) in the distribution—is also due to the choice of sources and target. We next discuss how the dependency decomposition and $\mathrm{I}_{\mathrm{dep}}$ shed new insight into mechanistic redundancy and nonholistic synergy.

There are two aspects of the PID that do not directly reflect properties of the joint distribution, but rather are determined by which variables are selected as sources and which the 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 collectively influence the target. The dependency decomposition and $\mathrm{I}_{\mathrm{dep}}$ make these phenomena explicit.

A. Source versus Mechanistic Redundancy

An interesting concept within the PID domain is that of *mechanistic redundancy* [12]. In its simplest form, this is existence of redundant information when the sources are independent. The AND distribution given in Table II 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. 5 exceeds edge quantity $b - i = c - h$. 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. (Section V B below draws out this implication.)

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	

AND	$\mathrm{I}_{\min}$	$\mathrm{I}_{\mathrm{proj}}$	$\mathrm{I}_{\mathrm{BROJA}}$	$\mathrm{I}_{\mathrm{ccs}}$	$\mathrm{I}_{\mathrm{dep}}$
01	1/2	1/2	1/2	0.29248	0.27043
0	0	0	0	0.20752	0.22957
1	0	0	0	0.20752	0.22957
$0 \cdot 1$	0.31128	0.31128	0.31128	0.10376	0.08170

TABLE II. AND distribution exemplifies both mechanistic redundancy and nonholistic synergy.

Mechanistic redundancy is closely tied to the concept of *target monotonicity* [23]:

(TM) $\mathrm{I}_\cap [X_0 \cdot X_1 \to Y] \geq \mathrm{I}_\cap [X_0 \cdot X_1 \to f(Y)]$. *(target monotonicity)*

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

1. $\mathrm{I}_\cap [X_0 \cdot X_1 \to (X_0X_1)] = 0$,
2. The possibility of mechanistic redundancy, or
3. Target monotonicity.

In effect, any given method of quantifying the PID cannot simultaneously assign zero redundancy to the "two-bit copy" distribution, allow mechanistic redundancy, and obey target monotonicity. $\mathrm{I}_{\mathrm{dep}}$ does not satisfy **(TM)**. Reference [23] demonstrated a general construction that maps a redundancy measure not satisfying **(TM)** to one that does, in the process violating property Item 1 above.

B. Holistic versus Nonholistic Synergy

A notion somewhat complementary to mechanistic redundancy is nonholistic synergy. Holistic synergy, or the third-order connected information [36], is the difference $\mathrm{H}[X_0X_1Y] - \mathrm{H}_{\mathrm{X_0X_1:X_0Y:X_1Y}}[X_0X_1Y]$. This quantifies the amount of structure in the distribution that is only constrained by the full triadic probability distribution, not any subsets of marginals. This quantity appears as the edge labeled m in Fig. 5. 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. 5. This quantity has a natural interpretation: how much does the

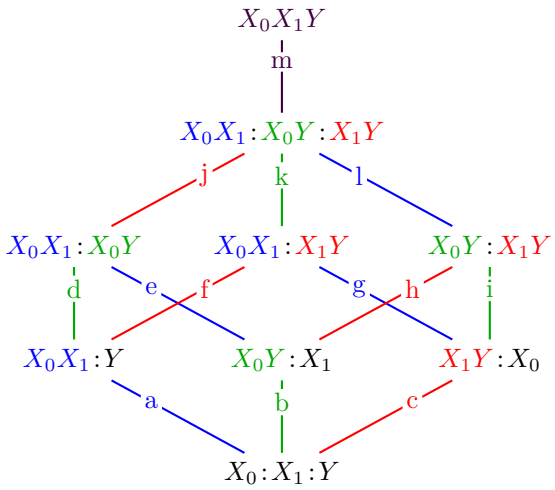

FIG. 5. Dependency structure for two source variables X_0 and X_1 and one target variable Y. Edges colored blue correspond to adding constraint X_0 X_1; edges colored green to adding constraint X_0 Y; and edges colored red to X_1 Y. The unique information $\mathrm{I}_{\mathrm{dep}}[X_0 \to Y]$ is calculated by considering the least change in $\mathrm{I}_\sigma[X_0X_1 : Y]$ along the green edges. See Appendix B and Fig. 7 for identities among the edges important for $\mathrm{I}_{\mathrm{dep}}$.

constraint X_0Y influence $\mathrm{I}[X_0X_1 : Y]$ in the context of the other dyadic relationships (X_0X_1, X_1Y), minus the unique information $\mathrm{I}_{\mathrm{dep}}[0 \to Y]$. The total PID synergy is then $\mathrm{I}_{\mathrm{dep}}[01 \to Y] = m + k - \min\{b, i, k\} = m + j - \min\{c, h, j\}$.

Here, again, the AND distribution seen in Table II exemplifies the phenomenon. The AND distribution is completely defined by the constraint $X_0X_1 : X_0Y : X_1Y$. That is, the AND distribution is the only distribution satisfying these pairwise constraints. This implies that the holistic synergy is zero. 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 co-information and therefore nonzero synergy and exhibits conditional dependence.

C. Shortcomings

$\mathrm{I}_{\mathrm{dep}}$ comes with its own concerns, however. First, it is defined using a minimum. Besides being mildly aesthetically displeasing, this can lead to nondifferentiable (though continuous) behavior, as seen in Fig. 4. Nondifferentiability can be seen as natural, as we have argued, if it coincides with a switch in qualitative behavior.

Perhaps more interestingly, $\mathrm{I}_{\mathrm{dep}}$ does not correspond to either the camel or elephant intuitions as described in Ref. [26] which proposes information-theoretical cryptography as a basis for unique information. In the relatively straightforward example of the PNT. UNQ. distribution, $\mathrm{I}_{\mathrm{dep}}$ does not correspond with any other proposed method of quantifying the PID. In this instance, it is simple to state why $\mathrm{I}_{\mathrm{dep}}$ quantifies the unique information as 1/4 bit: after constraining either X_0Y or X_1Y, constraining the other only increases $\mathrm{I}[X_0X_1 : Y]$ by 1/4 bit; that is, the unique value of either X_0Y or X_1Y is only a quarter because there exist contexts where that is all it can contribute to $\mathrm{I}[X_0X_1 : Y]$. However, it is difficult to see the operational meaning of this value. All other proposed methods match one or another secret key agreement rate. And so, they at least have a concrete operational interpretation.

Finally, $\mathrm{I}_{\mathrm{dep}}$ is a measure of unique information and so it cannot alone be used to quantify the PID with three or more sources. And, in the event where unique informations in concert with standard information measures are in fact sufficient for quantifying the entire PID, $\mathrm{I}_{\mathrm{dep}}$'s adherence to the identity axiom implies that it necessarily does not obey local positivity with three or more sources.

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 the Blackwell property **(BP)** and so, like $\mathrm{I}_{\mathrm{ccs}}$, it agrees with previous game-theoretic arguments raised in Ref. [17]. Unlike $\mathrm{I}_{\mathrm{ccs}}$, though, $\mathrm{I}_{\mathrm{dep}}$ satisfies **(LP)**. This makes it the only measure satisfying **(Id)** and **(LP)** which does not require 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. [18, 19, 22]. We agree with Ref. [22] that the likely solution is to employ 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, especially once the statistical significance of its structure applied to empirical data is explored.

ACKNOWLEDGMENTS

The authors thank Robin Ince and Daniel Feldspar 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] A. J. Gates and L. M. Rocha. Control of complex networks requires both structure and dynamics. *Scientific reports*, 6:24456, 2016. 1

[2] S. P. Faber, N. M. Timme, J. M. Beggs, and E. L. Newman. Computation is concentrated in rich clubs of local cortical neurons. *bioRxiv*, page 290981, 2018. 1

[3] V. Vijayaraghavan, R. G. James, and J. P. Crutchfield. Anatomy of a spin: the information-theoretic structure of classical spin systems. *Entropy*, 19(5):214, 2017. 1

[4] R. G. James, B. D. M. Ayala, B. Zakirov, and J. P. Crutchfield. Modes of information flow. *arXiv preprint arXiv:1808.06723*, 2018. 1

[5] R. B. Arellano-Valle, J. E. Contreras-Reyes, and M. G. Genton. Shannon entropy and mutual information for multivariate skew-elliptical distributions. *Scandinavian Journal of Statistics*, 40(1):42–62, 2013. 1

[6] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423, 623–656, 1948. 1

[7] C. E. Shannon. The bandwagon. *IEEE Transactions on Information Theory*, 2(3):3, 1956. 1

[8] C. E. Shannon. The lattice theory of information. *Trans. IRE Prof. Group Info. Th.*, 1(1):105–107, 1953. 1

[9] G. Birkhoff. *Lattice Theory*. American Mathematical Society, Providence, Rhode Island, first edition, 1940. 1, 2, 4

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

[11] R. G. James and J. P. Crutchfield. Multivariate dependence beyond shannon information. *Entropy*, 19(10):531, 2017. 1

[12] M. Harder, C. Salge, and D. Polani. Bivariate measure of redundant information. *Phys. Rev. E*, 87(1):012130, 2013. 1, 3, 8

[13] V. Griffith and C. Koch. Quantifying synergistic mutual information. In *Guided Self-Organization: Inception*, pages 159–190. Springer, 2014. 3, 6, 8

[14] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. 1, 3, 4, 7

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

[16] 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

[17] R. A.A. Ince. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy*, 19(7):318, 2017. 1, 4, 6, 9

[18] 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, 9

[19] 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. 9

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

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

[22] J. Rauh. Secret sharing and shared information. *Entropy*, 19(11):601, 2017. 1, 9

[23] J. Rauh, P. K. Banerjee, E. Olbrich, J. Jost, and N. Bertschinger. On extractable shared information. *Entropy*, 19(7):328, 2017. 1, 6, 8

[24] 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

[25] M. Zwick. An overview of reconstructability analysis. *Kybernetes*, 33(5/6):877–905, 2004. 1, 5, 6

[26] R. G. James, J. Emenheiser, and J. P. Crutchfield. A perspective on unique information: Directionality, intuitions, and secret key agreement. *arXiv:1808.08606*. 1, 4, 6, 9

[27] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006. 2, 5

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

[29] R. W. Yeung. *Information theory and network coding*. Springer, New York, 2008. 2

[30] E. W. Dijkstra. Why numbering should start at zero. 1982. 2

[31] 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

[32] U. M. Maurer and S. Wolf. Unconditionally secure key agreement and the intrinsic conditional information. *IEEE Transactions on Information Theory*, 45(2):499–514, 1999. 3

[33] 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

[34] N. Virgo and D. Polani. Decomposing multivariate information (abstract). 2017. Accessed 26 April 2018, `https://www.mis.mpg.de/fileadmin/pdf/abstract_gso18_3302.pdf`. 5

[35] S. Amari. Information geometry on hierarchy of probability distributions. *IEEE transactions on information theory*, 47(5):1701–1711, 2001. 5

[36] E. Schneidman, S. Still, M. J. Berry, and W. Bialek. Network information and connected correlations. *Phys. Rev. Lett.*, 91(23):238701, 2003. 5, 8, 13

[37] J. Runge. Quantifying information transfer and mediation along causal pathways in complex systems. *Phys. Rev. E*, 92(6):062829, 2015. 5

[38] J. Sun and E. M. Bollt. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D: Nonlinear Phenomena*, 267:49–57, 2014. 5

[39] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, New York, 1988. 5

[40] R. G. James, C. J. Ellison, and J. P. Crutchfield. dit: a Python package for discrete information theory. *The Journal of Open Source Software*, 3(25):738, 2018. 6, 7

[41] C. Finn and J. T. Lizier. Pointwise partial information decomposition using the specificity and ambiguity lattices. *Entropy*, 20(4):297, 2018. 6

Appendix A: Constrained Three-Variable Maximum Entropy Distributions

Here, we characterize the maximum entropy distributions defined in Eq. (6) and used to populate the dependency structure. In particular, we give maximum entropy distributions for forms of marginal constraints that describe the lowest three levels of the constraint lattice as shown in Fig. 2.

Let us fix notation. Consider a joint distribution $p(ABC)$ and maximum entropy distributions for some constraint set σ: $p_\sigma(ABC)$. We label information measures and other quantities that are computed relative to this constrained maximum entropy distribution with the subscript σ: for example, $\mathrm{H}_{\mathrm{A:B:C}}[ABC]$ refers to the entropy of the product distribution $p(abc) = p(a)p(b)p(c)$, as this is the distribution consistent with the constraint $A : B : C$ and has maximum entropy. Such quantities without a subscript are calculated from the original distribution. In our use of the dependency structure to quantify unique information, we are interested in $\mathrm{I}_\sigma[AB{:}C]$ as this will represent the sources-target mutual information.

The lowest node in the dependency lattice constrains all single-variable marginal distributions, but no pairwise marginal distributions. With only single-variable marginal distributions constrained, the maximum entropy distribution is such that the variables are independent, also known as the product distribution:

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

It can be seen that this must be the maximum entropy distribution, since an increase in any mutual information corresponds to an equal decrease in at least two conditional entropies, resulting in a lower total entropy. The informational structure of this distribution can be seen in Fig. 6a.

The first row up in the constraint lattice captures those antichains that contain one pairwise constraint and one single-variable constraint. The maximum entropy distribution corresponding to this constraint set is given by:

$$p_{AC:B}(ABC) = p(AC)p(B)\ . \tag{A2}$$

All atoms of the information diagram that capture the overlap of $\mathrm{H}[AC]$ and $\mathrm{H}[B]$ vanish. Again, it can be seen that the maximum entropy distribution must take this form, since any deviation from the information partitioning seen in Fig. 6b, which satisfies the constraint $AC : B$, must result in an overall decrease to the entropy.

The elements in the second row of the constraint lattice include constraints on two pairwise marginal distributions each. These constraints both contain one of the variables, and the maximum entropy distribution takes on the following Markov form:

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

This distribution forms a Markov chain A–⟜–C–⟜–B and therefore $\mathrm{I}_{\mathrm{AC:BC}}[A{:}B|C] = 0$. To see that this must be the form of the maximum entropy distribution, consider the expansion:

$$\mathrm{H}[ABC] = \mathrm{H}[C] + \mathrm{H}[A|C] + \mathrm{H}[B|C] - \mathrm{I}[A{:}B|C]\ .$$

The first three terms of the righthand side are constrained by $p(C)$, $p(AC)$, and $p(BC)$, respectively. Since the conditional mutual information is necessarily nonnegative, the final term being zero corresponds to the distribution with the maximum entropy. Such a distribution is realized by the given Markov chain. The mutual information $\mathrm{I}_{\mathrm{AC:BC}}[AB{:}C]$ is equal to:

$$\mathrm{I}[A:C] + \sum_{\substack{a\in\mathcal{A}\\ b\in\mathcal{B}\\ c\in\mathcal{C}}} p(a,b,c)\log_2\frac{p(a)p(b,c)}{p(c)p(a,b)}\ .$$

The information diagrams for each of these three distributions are given in Fig. 6.

With the structure of these distributions in hand, we now

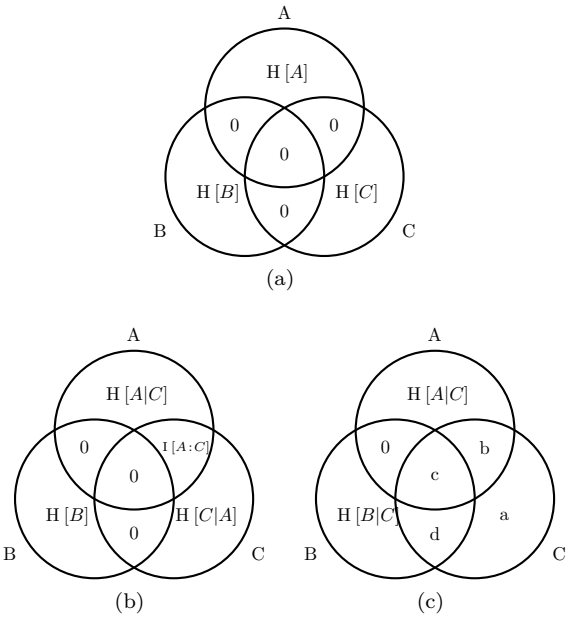

FIG. 6. Information diagrams corresponding to the maximum entropy distributions described in Eqs. (A1) to (A3). The four variables in subfigure (c) satisfy $a+b+c+d = \mathrm{H}[C]$, $b+c = \mathrm{I}[A{:}C]$, and $c+d = \mathrm{I}[B{:}C]$.

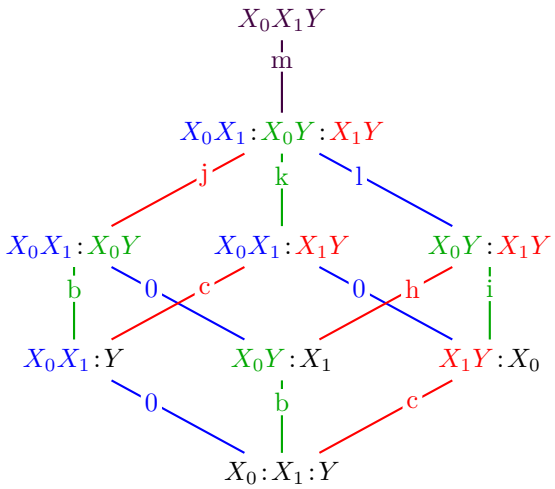

FIG. 7. The reduced form of the dependency lattice quantified by sources-target mutual information, with $b = \mathrm{I}[X_0 : Y]$ and $c = \mathrm{I}[X_1 : Y]$. All edges are guaranteed to be nonnegative except for l.

turn to proving several properties of the $\mathrm{I}_{\mathrm{dep}}$ measure.

Appendix B: Sources-Target Dependency Structure

Interpreting the set of antichain covers as possible marginal constraints on probability distributions, we defined a dependency lattice that gradually introduces dependencies into an otherwise unstructured distribution. In this method of quantifying the partial information decomposition, $\mathrm{I}_{\mathrm{dep}}$ is defined according to the node-node differences of the sources-target mutual information $\mathrm{I}[X_0X_1 : Y]$. These lattice edges are labeled in Fig. 5.

The maximum entropy distributions on the lowest three levels of nodes follow the forms given in Appendix A. By appropriately assigning X_0, X_1, and Y to A, B, and C, we obtain the relationships arising among the edges summarized in Fig. 7 and the following results for the sources-target mutual information:

$$\mathrm{I}_{\mathrm{X_0:X_1:Y}}[X_0X_1 : Y] = 0 \tag{B1}$$
$$\mathrm{I}_{\mathrm{X_0X_1:Y}}[X_0X_1 : Y] = 0 \tag{B2}$$
$$\mathrm{I}_{\mathrm{X_0Y:X_1}}[X_0X_1 : Y] = \mathrm{I}[X_0 : Y] \tag{B3}$$
$$\mathrm{I}_{\mathrm{X_1Y:X_0}}[X_0X_1 : Y] = \mathrm{I}[X_1 : Y] \tag{B4}$$
$$\mathrm{I}_{\mathrm{X_0X_1:X_0Y}}[X_0X_1 : Y] = \mathrm{I}[X_0 : Y] \tag{B5}$$
$$\mathrm{I}_{\mathrm{X_0Y:X_1Y}}[X_0X_1 : Y] = \mathrm{I}[X_1 : Y] \tag{B6}$$

Now, we can simply read off the values of a through g, and derive other simple relationships:

$$\begin{aligned} a = e = g &= 0 \\ d = b &= \mathrm{I}[X_0 : Y] \\ f = c &= \mathrm{I}[X_1 : Y] \\ b + h &= c + i \\ b + j + m = c + k + m &= \mathrm{I}[X_0X_1 : Y] \ . \end{aligned}$$

Furthermore:

$$\begin{aligned} h &= \mathrm{I}_{\mathrm{X_0Y:X_1Y}}[X_0X_1 : Y] - \mathrm{I}[X_0 : Y] \\ &= \mathrm{I}_{\mathrm{X_0Y:X_1Y}}[X_1 : Y | X_0] \\ i &= \mathrm{I}_{\mathrm{X_0Y:X_1Y}}[X_0X_1 : Y] - \mathrm{I}[X_1 : Y] \\ &= \mathrm{I}_{\mathrm{X_1Y:X_1Y}}[X_0 : Y | X_1] \\ j &= \mathrm{I}_{\mathrm{X_0X_1:X_0Y:X_1Y}}[X_0X_1 : Y] - \mathrm{I}[X_0 : Y] \\ &= \mathrm{I}_{\mathrm{X_0X_1:X_0Y:X_1Y}}[X_1 : Y | X_0] \\ k &= \mathrm{I}_{\mathrm{X_0X_1:X_0Y:X_1Y}}[X_0X_1 : Y] - \mathrm{I}[X_1 : Y] \\ &= \mathrm{I}_{\mathrm{X_0X_1:X_1Y:X_1Y}}[X_0 : Y | X_1] \ . \end{aligned}$$

Our first task is to demonstrate that all edges, save the one labeled l, correspond to nonnegative differences in the sources-target mutual information. The edges b and c are given by mutual informations and so must be nonnegative. The edges h, i, j, and k are given by conditional mutual informations computed relative to certain maximum entropy distributions and, therefore, must also be nonnegative. The edge labeled m is the third-order connected information [36], which can be written as a Kullback-Leibler divergence and so must also be nonnegative. This leaves only the edge l potentially negative. These last two edges, l and m, do not involve the addition of a source-target constraint and so are not considered in computing $\mathrm{I}_{\mathrm{dep}}$.

Next, we demonstrate that only two edges meaningfully contribute to the determination of $\mathrm{I}_{\mathrm{dep}}$. Since the coinformation $\mathrm{I}_{\mathrm{X_0Y:X_1Y}}[X_0 : X_1 : Y] = \mathrm{I}_{\mathrm{X_0Y:X_1Y}}[X_0 : X_1]$ is necessarily nonnegative, we find that:

$$\begin{aligned} h &\leq \mathrm{I}[X_1 : Y] = c \\ i &\leq \mathrm{I}[X_0 : Y] = b \ . \end{aligned}$$

And so, in computing $\mathrm{I}_{\mathrm{dep}}$ one need only consider the edges i and k (for $\mathrm{I}_{\mathrm{dep}}[X_0 \to Y]$) or h and j (for $\mathrm{I}_{\mathrm{dep}}[X_1 \to Y]$).

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

This section establishes the properties of the bivariate partial information decomposition induced by $\mathrm{I}_{\mathrm{dep}}$.

Self-redundancy

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

The shared information for an antichain with a single subset of source variables is precisely the mutual information between those source variables and the target.

We take this axiom constructively, defining three of the four shared informations $\mathrm{I}_\cap$ and providing three constraints on partial informations I_∂ in Eqs. (2) to (4).

Nonnegativity

Property **(LP)**: For all antichains σ: $\mathrm{I}_\partial[\sigma \to Y] \geq 0$.

Every partial information value resulting from the mobius inversion of the redundancy lattice is nonnegative.

We begin with the unique partial information from X_0. Both arguments of the minimum in $\mathrm{I}_{\mathrm{dep}}[0 \to Y]$ were shown to be nonnegative in Appendix B:

$$\mathrm{I}_\partial[0 \to Y] = \mathrm{I}_{\mathrm{dep}}[0 \to Y] = \min(i, k) \geq 0 \ .$$

Using the self-redundancy axiom to define $\mathrm{I}_\cap[0 \to Y] = \mathrm{I}[X_0 : Y]$ and knowing that $\mathrm{I}_\partial[0 \to Y] = \min(i, k) \leq \mathrm{I}[X_0 : Y]$ from Appendix B, we have:

$$\mathrm{I}_\partial[0 \cdot 1 \to Y] = \mathrm{I}_\cap[0 \to Y] - \mathrm{I}_\partial[0 \to Y] \geq 0 \ .$$

To determine the signs of the remaining two partial information atoms, we must consider the ordering of i and k.

Reductions are done by using results of Appendix B. We repeatedly use the redundancy lattice inversions: $\mathrm{I}_\partial[1 \to Y] = \mathrm{I}_\cap[1 \to Y] - \mathrm{I}_\partial[0 \cdot 1 \to Y]$ and $\mathrm{I}_\partial[01 \to Y] = \mathrm{I}_\cap[01 \to Y] - \mathrm{I}_\partial[0 \cdot 1 \to Y] - \mathrm{I}_\partial[0 \to Y] - \mathrm{I}_\partial[1 \to Y]$.

CASE 1: $i \leq k$.

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

CASE 2: $k \leq i$.

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

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

Monotonicity

Property **(M)**: $\alpha \preceq \beta \implies \mathrm{I}_\cap[\alpha \to Y] \leq \mathrm{I}_\cap[\beta \to Y]$.

For any two antichains α, β, an ordering between them implies the same ordering of their shared informations.

This follows immediately from **(LP)** above.

Symmetry

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

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

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 $\mathrm{I}_{\mathrm{dep}}[0 \to Y]$ or $\mathrm{I}_{\mathrm{dep}}[1 \to Y]$.

Computing $\mathrm{I}_{\mathrm{dep}}[0 \to Y] = \min(b, i, k)$ gives $\mathrm{I}_\partial[1 \to Y] = \min(c, h, j)$, 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 Appendix C cases, $\mathrm{I}_\partial[1 \to Y]$ 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)**:

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

Consider sources X_0 and X_1 and output $Y = X_0X_1$, the concatenation of inputs. The mutual information of either source with the target is simply the entropy of that source. That is, $b = \mathrm{H}[X_0]$. Using appropriate permutations of Eqs. (A2) and (A3) ($A = X_0$, $B = Y$, $C = X_1$), we find that $i = \mathrm{H}[X_0|X_1]$. Now, starting at the constraint $\sigma = X_0X_1 : X_1Y$ as in Eq. (A3) ($A = X_0$, $B = X_1$, $C = Y$), we see that additionally constraining $p(X_0Y)$ fully constrains the distribution to its original form, with a sources-target mutual information of $\mathrm{H}[X_0X_1]$. That is, $k = \mathrm{H}[X_0|X_1]$. The minimum of these three quantities gives $\mathrm{I}_{\mathrm{dep}}[0 \to Y] = \mathrm{H}[X_0|X_1]$ and therefore verifies the identity axiom.