

Decomposing Multivariate Information

Paul L. Williams and Randall D. Beer

Abstract—Mutual information provides a general tool for quantifying the dependency between two variables, but its generalization to multivariate interactions has long been problematic. Here we develop from first principles a novel solution to the problem of quantifying multivariate information. We begin with an axiomatic definition of redundancy as the overlapping information that several sources share about a given variable. We then show how this definition of redundancy induces a lattice over sets of sources that clarifies the general structure of multivariate information. From this lattice, we derive a definition of partial information atoms that exhaustively decompose the Shannon information in a multivariate system in terms of the redundancy between synergies of subsets of the sources. Finally, we discuss the relationship between our partial information decomposition and interaction information, the current de facto measure of multivariate information. Unlike interaction information, the atoms of our partial information decomposition are never negative and always support a clear interpretation as informational quantities. Our analysis also demonstrates how the confusing negativity of interaction information can be explained by its confounding of redundancy and synergy.

Index Terms—interaction information, multivariate mutual information, redundancy, synergy, information diagrams

I. INTRODUCTION

FROM its roots in Shannon’s seminal work on reliability and coding in communication systems, information theory has grown into a ubiquitous general tool for the analysis of complex systems, with applications in neuroscience, genetics, physics, machine learning, and many other areas. Perhaps the most widely used concept from information theory is Shannon’s mutual information [1], [2], which provides a general measure of the interdependence between two variables, or two sets of variables. Given this general applicability, there has also been great interest in extending mutual information to multivariate interactions, but despite numerous attempts [3]–[10] this remains largely an open problem. Meanwhile, many of the most interesting and challenging scientific questions, such as many-body problems in physics, n -person games in game theory, and population coding in neuroscience, involve understanding the structure of interactions between three or more variables.

The two main attempts to generalize mutual information to multivariate interactions are the *total correlation* proposed by Watanabe [4] and the *interaction information* of McGill [3]¹. The total correlation, as its name suggests, measures the total

dependency between a set of variables as a single monolithic quantity. Thus, the total correlation does not provide any insight into how dependencies are distributed amongst the variables, i.e., it says nothing about the *structure* of multivariate information.

In contrast, interaction information was proposed as a measure of the amount of information bound up in a set of variables beyond that which is present in any subset of those variables. Thus, entropy and mutual information correspond to first- and second-order interaction information, respectively, and together with its higher-order variants, interaction information provides a way of characterizing the structure of multivariate information. Interaction information is also the natural generalization of mutual information when Shannon entropy is viewed as a signed measure on information diagrams [2], [13], [15]. However, the wider use of interaction information has largely been hampered by the “odd” [13] and “unfortunate” [2] property that, for three or more variables, it can be negative (see also [7], [15]–[18]). For information as it is commonly understood, it is entirely unclear what it means for one variable to provide “negative information” about another. Moreover, as we will show, this confusing property of negativity is actually symptomatic of deeper problems regarding the interpretation of interaction information for higher-dimensional systems.

Here we formulate a new perspective on the structure of multivariate information. Beginning from first principles, we propose a definition of redundancy that captures the overlapping information that two or more sources share about a given variable. Then we show how this definition of redundancy forms a lattice over the set of possible information sources that clarifies the general structure of multivariate information. Next we use this definition to exhaustively decompose the Shannon information in a multivariate system into partial information atoms consisting of redundancies between synergies of subsets of the sources. Unlike interaction information, the atoms of our partial information decomposition are never negative and always support a clear interpretation as informational quantities. Finally, our analysis also demonstrates how the negativity of interaction information can be explained by its confounding of redundant and synergistic interactions.

II. THE STRUCTURE OF MULTIVARIATE INFORMATION

Suppose we are given a random variable S and a random vector $\mathbf{R} = \{R_1, R_2, \dots, R_{n-1}\}$. Then our goal is to decompose the information that $\mathbf{R}$ provides about S in terms of the partial information contributed either individually or jointly by various subsets of $\mathbf{R}$. For example, in a neuroscience context, S may correspond to a stimulus that takes on different values and $\mathbf{R}$ to the evoked responses of different neurons. In this case, we would like to quantify the information that the joint

Paul L. Williams is with the Cognitive Science Program and Randall D. Beer is with the Cognitive Science Program and the School of Informatics and Computing, Indiana University, Bloomington, IN, 47406 USA. This work was supported in part by NSF grant IIS-0916409 to RDB and an NSF IGERT traineeship to PLW.

¹Total correlation is also known as the multivariate constraint [11], multiinformation [8], and integration [12]. Interaction information is also known as multiple mutual information [7], co-information [13], and synergy [14].

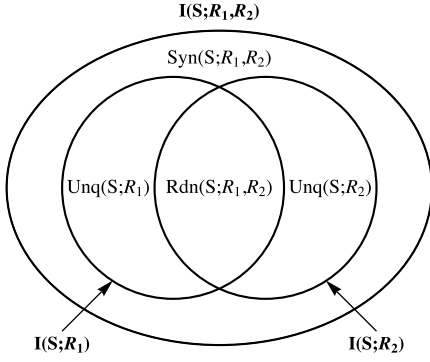


Fig. 1. Structure of information for 3 variables. Regions correspond to unique information (*Unq*), redundancy (*Rdn*), and synergy (*Syn*).

neural response provides about the stimulus, and to distinguish between information due to responses of individual neurons versus combinations of them [14], [19].

Consider the simplest case of a system with three variables. How much total information does $\mathbf{R} = \{R_1, R_2\}$ provide about S ? How do R_1 and R_2 contribute to the total information? The answer to the first question is given by the mutual information (MI) $I(S; R_1, R_2)$, while for the latter we can identify three distinct possibilities. First, R_1 may provide information that R_2 does not, or vice versa (*unique information*). For example, if R_1 is a copy of S and R_2 is a degenerate random variable, then the total information from $\mathbf{R}$ reduces to the unique information from R_1 . Second, R_1 and R_2 may provide the same or overlapping information (*redundancy*). For example, if R_1 and R_2 are both copies of S then they redundantly provide complete information. Third, the combination of R_1 and R_2 may provide information that is not available from either alone (*synergy*). A well-known example for binary variables is the exclusive-OR function $S = R_1 \oplus R_2$, in which case R_1 and R_2 individually provide no information but together provide complete information. Thus, intuitively, the total information from $\mathbf{R}$ decomposes into unique information from R_1 and R_2 , redundant information shared by R_1 and R_2 , and synergistic information contributed jointly by R_1 and R_2 (Fig. 1).

In sum, for three variables we can identify unique information, redundancy, and synergy as the basic atoms of multivariate information. In fact, as later developments will clarify, unique information is best thought of as a degenerate form of redundancy or synergy, so that redundancy and synergy alone constitute the basic building blocks of multivariate information. In particular, we will find that combinations of redundancy and synergy, which may at first sound paradoxical, play a fundamental role in structuring multivariate information in higher dimensions. Next we proceed to formalize these ideas, beginning with the notion of redundancy.

III. REDUNDANCY AS OVERLAPPING INFORMATION

Let $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_k$ be potentially overlapping subsets of $\mathbf{R}$, called *sources*. How can we quantify the redundant information that all sources provide about S ? The intuition that we want to capture is that redundancy corresponds to

overlapping or shared information. In particular, redundancy should be analogous to intersection from set theory, measuring the common information “contained” by two or more sources.

To capture this idea of intersecting information, we define redundancy as a $(k + 1)$ -argument function

$$I_{\cap} : S \times \underbrace{2^{\mathbf{R}} \times \dots \times 2^{\mathbf{R}}}_k \mapsto \mathbb{R},$$

$$(S, \mathbf{A}_1, \dots, \mathbf{A}_k) \mapsto I_{\cap}(S; \mathbf{A}_1, \dots, \mathbf{A}_k) \quad (1)$$

that satisfies the following axioms:

Axiom 1 (symmetry). $I_{\cap}$ is symmetric in the $\mathbf{A}_i$ ’s.

Axiom 2 (self-redundancy). $I_{\cap}(S; \mathbf{A}) = I(S; \mathbf{A})$.

Axiom 3 (monotonicity). $I_{\cap}(S; \mathbf{A}_1, \dots, \mathbf{A}_{k-1}, \mathbf{A}_k) \leq I_{\cap}(S; \mathbf{A}_1, \dots, \mathbf{A}_{k-1})$ with equality if $\mathbf{A}_{k-1} \subseteq \mathbf{A}_k$.

Note that these axioms are analogous to basic properties of set intersection; namely, that it is commutative (symmetric):

$$X \cap Y = Y \cap X,$$

idempotent (‘self-intersection’ is preserving):

$$X \cap X = X,$$

and monotonic:

$$(X_1 \cap \dots \cap X_{k-1} \cap X_k) \subseteq (X_1 \cap \dots \cap X_{k-1})$$

with equality if $X_{k-1} \subseteq X_k$. Thus, we claim that any measure of redundancy as intersecting information must satisfy this axiomatic definition. Two useful properties follow immediately from these axioms.

Corollary 1. $I_{\cap}$ is nonnegative.

Proof: Since $\emptyset \subseteq \mathbf{A}_i$ for any $\mathbf{A}_i$,

$$I_{\cap}(S; \mathbf{A}_1, \dots, \mathbf{A}_k, \emptyset) = I_{\cap}(S; \emptyset) = I(S; \emptyset) = 0$$

by Axs. 2 and 3. Also from Ax. 3, we have that

$$I_{\cap}(S; \mathbf{A}_1, \dots, \mathbf{A}_k) \geq I_{\cap}(S; \mathbf{A}_1, \dots, \mathbf{A}_k, \emptyset),$$

so we conclude that $I_{\cap}(S; \mathbf{A}_1, \dots, \mathbf{A}_k) \geq 0$. $\square$

Corollary 2. $I(A; O_i) = I_{\cap}(A; O_i) \geq I_{\cap}(A; O_i, \dots)$.

Thus, redundancy is nonnegative, bounded from above by the information from any one source, and maximized by the self-redundancy, analogous to the property that MI is maximized by the self-information $I(S; S) = H(S)$.

An appealing candidate measure of redundancy that satisfies the axiomatic definition $I_{\cap}$ is

$$I_{\min}(S; \mathbf{A}_1, \dots, \mathbf{A}_k) = \sum_s p(s) \min_{\mathbf{A}_i} I(S = s; \mathbf{A}_i) \quad (2)$$

where

$$I(S = s; \mathbf{A}) = \sum_{\mathbf{a}} p(\mathbf{a}|s) \left[\log \frac{1}{p(s)} - \log \frac{1}{p(s|\mathbf{a})} \right] \quad (3)$$

is the *specific information* that $\mathbf{A}$ provides about each outcome $s \in S$ (see also Appendix A). The term $\frac{1}{p(s)}$ is the surprise of

s [2], so $I(S = s; \mathbf{A})$ is the average reduction in surprise of s given knowledge of $\mathbf{A}$. Note also that

$$I(S; \mathbf{A}) = \sum_s p(s) I(S = s; \mathbf{A}), \quad (4)$$

so that both I and $I_{\min}$ are expected values of specific information terms. Thus, $I_{\min}$ quantifies redundancy as the minimum information that any source provides about each outcome of S , averaged over all possible outcomes. This definition captures the idea that redundancy is the information shared by all sources (the minimum that any one provides), while taking into account that sources may provide information about different outcomes of S .

Theorem 1. $I_{\min}$ satisfies the axiomatic definition $I_{\cap}$.

The proof of this theorem is in Appendix E, as are all proofs not included in the main text. Henceforth, we are careful to use $I_{\cap}$ when discussing properties that follow from the axiomatic definition, and $I_{\min}$ when discussing properties specific to that measure. $I_{\min}$ has the following additional desirable properties.

Theorem 2. $I_{\min}(S; \mathbf{A}_1, \dots, \mathbf{A}_k) = 0$ if, and only if, for every joint outcome $(s, \mathbf{a}_1, \dots, \mathbf{a}_k)$,

$$p(s, \mathbf{a}_i) = p(s)p(\mathbf{a}_i)$$

for at least one $\mathbf{a}_i$.

Theorem 2 is a natural generalization of the property that $I(S; \mathbf{A}) = 0$ if, and only if, for every joint outcome $(s, \mathbf{a})$,

$$p(s, \mathbf{a}) = p(s)p(\mathbf{a}).$$

Thus, according to this theorem, redundancy (defined as $I_{\min}$) is zero if, and only if, for each joint outcome one or more of the sources provide no information about S .

Theorem 3. If $S \rightarrow \mathbf{A}_k \rightarrow \mathbf{A}_{k-1}$, then

$$I_{\min}(S; \mathbf{A}_1, \dots, \mathbf{A}_{k-1}, \mathbf{A}_k) = I_{\min}(S; \mathbf{A}_1, \dots, \mathbf{A}_{k-1}).$$

In words, Theorem 3 states that if S , $\mathbf{A}_k$, and $\mathbf{A}_{k-1}$ form a Markov chain in that order (i.e., if $\mathbf{A}_k$ screens off $\mathbf{A}_{k-1}$ from S), then any information provided by $\mathbf{A}_{k-1}$ is also provided redundantly by $\mathbf{A}_k$. This theorem generalizes the data processing inequality [2], as the following corollary shows:

Corollary 3 (data processing inequality). If $S \rightarrow \mathbf{A}_i \rightarrow \mathbf{A}_j$, then $I(S; \mathbf{A}_i) \geq I(S; \mathbf{A}_j)$.

Proof: If $S \rightarrow \mathbf{A}_i \rightarrow \mathbf{A}_j$, then

$$I(S; \mathbf{A}_j) = I_{\min}(S; \mathbf{A}_j) = I_{\min}(S; \mathbf{A}_i, \mathbf{A}_j).$$

We also know that

$$I(S; \mathbf{A}_i) = I_{\min}(S; \mathbf{A}_i) \geq I_{\min}(S; \mathbf{A}_i, \mathbf{A}_j),$$

so we conclude that $I(S; \mathbf{A}_i) \geq I(S; \mathbf{A}_j)$. $\square$

The following corollary states that complete dependence (between sources) implies complete redundancy.

Corollary 4. If $\mathbf{A}_j = f(\mathbf{A}_i)$, then

$$I_{\min}(S; \mathbf{A}_i, \mathbf{A}_j) = I(S; \mathbf{A}_j).$$

Proof: If $\mathbf{A}_j = f(\mathbf{A}_i)$, then $S \rightarrow \mathbf{A}_i \rightarrow \mathbf{A}_j$. $\square$

However, it is worth noting that, aside from this connection, dependence between sources and redundancy are completely dissociable: neither does dependence imply redundancy, nor the converse².

What are the distinct ways in which collections of sources might contribute redundant information? Formally, answering this question means identifying the domain of $I_{\cap}$. Thus far, we have assumed that the natural domain is the collection of all possible sets of sources, but in fact this domain can be greatly simplified. Given sources $\mathbf{A}_i$ and $\mathbf{A}_j$ with $\mathbf{A}_i \subseteq \mathbf{A}_j$, Axiom 3 states that the redundancy for $\mathbf{A}_i$ and $\mathbf{A}_j$ is equivalent to the redundancy when $\mathbf{A}_j$ is removed:

$$I_{\cap}(S; \mathbf{A}_i, \mathbf{A}_j, \dots) = I_{\cap}(S; \mathbf{A}_i, \dots).$$

Applying this property recursively, it follows that for any collection of sources where some are supersets of others, the redundancy for that collection is equivalent to the redundancy when all supersets are removed. Thus, the domain for $I_{\cap}$ can be reduced to the collection of all sets of sources such that no source is a superset of any other:

$$\mathcal{A}(\mathbf{R}) = \{\alpha \in \mathcal{P}_1(\mathcal{P}_1(\mathbf{R})) : \forall \mathbf{A}_i, \mathbf{A}_j \in \alpha, \mathbf{A}_i \not\subseteq \mathbf{A}_j\}, \quad (5)$$

where $\mathcal{P}_1(\mathbf{R}) = \mathcal{P}(\mathbf{R}) \setminus \{\emptyset\}$ is the set of all nonempty subsets of $\mathbf{R}$. Formally, $\mathcal{A}(\mathbf{R})$ corresponds to the set of all *antichains* on the inclusion lattice $\langle \mathcal{P}(\mathbf{R}), \subseteq \rangle$, excluding the empty set³. Henceforth, we will denote elements of $\mathcal{A}(\mathbf{R})$, corresponding to collections of sources, with bracketed expressions containing only the indices for each source. For instance, $\{\{R_1, R_2\}\}$ will be $\{12\}$, $\{\{R_1\}, \{R_2, R_3\}\}$ will be $\{1\}\{23\}$, and so forth.

The possibilities for redundancy are also naturally structured, which can be shown by extending the same line of reasoning to define a partial ordering on the elements of $\mathcal{A}(\mathbf{R})$. Consider two collections of sources, $\alpha, \beta \in \mathcal{A}(\mathbf{R})$, where for each source $\mathbf{B} \in \beta$ there exists a source $\mathbf{A} \in \alpha$ such that $\mathbf{A} \subseteq \mathbf{B}$. This means that for each $\mathbf{B} \in \beta$ there is an $\mathbf{A} \in \alpha$ such that $\mathbf{A}$ provides no more information than $\mathbf{B}$. The redundant information shared by all $\mathbf{B} \in \beta$ must therefore at least include any redundant information shared by all $\mathbf{A} \in \alpha$. Thus, we can define a partial order $\preceq$ over the elements of $\mathcal{A}(\mathbf{R})$ such that one element (collection of sources) precedes another if, and only if, the latter provides any redundant information that the former provides:

$$\forall \alpha, \beta \in \mathcal{A}(\mathbf{R}), (\alpha \preceq \beta \Leftrightarrow \forall \mathbf{B} \in \beta, \exists \mathbf{A} \in \alpha, \mathbf{A} \subseteq \mathbf{B}). \quad (6)$$

Applying this ordering to the elements of $\mathcal{A}(\mathbf{R})$ produces a *redundancy lattice*⁴, in which a higher element provides at least as much redundant information as a lower one (Fig. 2).

From the redundancy lattice, it is possible to read off some of the properties of $I_{\cap}$ noted earlier. For instance, the property that redundancy is maximized by self-redundancy corresponds to the fact that any node representing a single source is higher in the lattice than any other node involving that source. For example, in Fig. 2B, the node labeled $\{12\}$, corresponding

²The redundancy/dependence relationship is discussed in Appendix B.

³Basic concepts from lattice theory are reviewed in Appendix C.

⁴Details regarding the redundancy lattice are given in Appendix D.

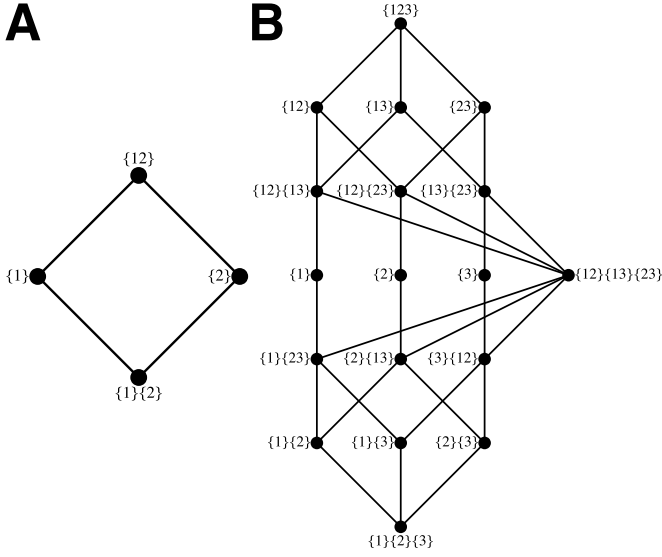


Fig. 2. Redundancy lattice for (A) 3 and (B) 4 variables.

to self-redundancy for the source $\{R_1, R_2\}$, is higher than the nodes labeled $\{12\}\{13\}$, $\{12\}\{13\}\{23\}$, and $\{3\}\{12\}$. Another useful property of $I_\cap$ relates to the top and bottom nodes of the lattice. The top node corresponds to the self-redundancy for $\mathbf{R}$, reflecting the fact that $I_\cap$ is bounded from above by the total information from $\mathbf{R}$. At the other end of the spectrum, the bottom element corresponds to the redundant information that each individual variable in $\mathbf{R}$ provides, with all other possibilities for redundancy falling between these two extremes.

IV. PARTIAL INFORMATION DECOMPOSITION

The redundant information associated with each node of the redundancy lattice includes, but is not limited to, the redundant information provided by all nodes lower in the lattice. Thus, moving from node to node up the lattice, $I_\cap$ can be thought of as a kind of “cumulative information function,” effectively integrating the information provided by increasingly inclusive collections of sources. Next, we define an inverse of $I_\cap$ called the *partial information (PI) function*. Whereas $I_\cap$ quantifies cumulative information, the PI-function measures the partial information contributed uniquely by each particular collection of sources. This partial information will form the atoms into which we decompose the total information that $\mathbf{R}$ provides about S .

For a collection of sources $\alpha \in \mathcal{A}(\mathbf{R})$, the PI-function, denoted I_∂ , corresponds to the Möbius inverse of $I_\cap$ [20], [21] and is defined implicitly by

$$I_\cap(S; \alpha) = \sum_{\beta \preceq \alpha} I_\partial(S; \beta). \quad (7)$$

From this relationship, it is clear that I_∂ can be calculated recursively as

$$I_\partial(S; \alpha) = I_\cap(S; \alpha) - \sum_{\beta \prec \alpha} I_\partial(S; \beta). \quad (8)$$

Thus, in words, $I_\partial(S; \alpha)$ quantifies the information provided redundantly by the sources of α that is not provided by any

simpler collection of sources (i.e., any β lower than α on the redundancy lattice). The following theorem provides a closed-form expression for I_∂ based on the inclusion-exclusion principle:

Theorem 4.

$$I_\partial(S; \alpha) = I_\cap(S; \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathcal{B} \subseteq \alpha^- \\ |\mathcal{B}|=k}} I_\cap(S; \bigwedge \mathcal{B})$$

where $\bigwedge$ is the meet operator and α^- represents the nodes immediately below α on the redundancy lattice.

The PI-function is nonnegative if, and only if, $I_\cap$ is *totally monotone*⁵ [25], [26], meaning that

$$I_\cap(S; \bigvee_{1 \leq j \leq k} \alpha_j) \geq \sum_{\emptyset \neq J \subseteq \{1, \dots, k\}} (-1)^{|J|-1} I_\cap(S; \bigwedge_{j \in J} \alpha_j) \quad (9)$$

for $k \geq 2$ and every k -element subset of $\mathcal{A}(\mathbf{R})$. Total monotonicity is a stronger form of supermodularity⁶, which, along with the dual concept of submodularity, has recently been discussed as a central or even defining property of information [27]–[30]. For example, in [29] submodularity is incorporated into an axiomatic definition of entropy, and it is shown that (conditional) MI is nonnegative if, and only if, entropy is submodular. Thus, there is an attractive parallel between the relationships (nonnegative MI $\leftrightarrow$ submodular entropy) and (nonnegative PI $\leftrightarrow$ totally monotone redundancy). Later, we will show that $I_{\min}$ has the desired property of total monotonicity, so that its corresponding PI-function decomposes $I(S; \mathbf{R})$ into a sum of nonnegative PI-terms.

The decomposition of $I(S; \mathbf{R})$ into a sum of PI-terms follows from

$$I(S; \mathbf{R}) = I_\cap(S; \mathbf{R}) = \sum_{\beta \preceq \{\mathbf{R}\}} I_\partial(S; \beta). \quad (10)$$

For the 3-variable case $\mathbf{R} = \{R_1, R_2\}$, Equation (10) yields

$$I(S; R_1) = I_\partial(S; \{1\}) + I_\partial(S; \{1\}\{2\}) \quad (11)$$

and

$$I(S; R_1, R_2) = I_\partial(S; \{1\}) + I_\partial(S; \{2\}) + I_\partial(S; \{1\}\{2\}) + I_\partial(S; \{12\}). \quad (12)$$

The relationship between these equations can be represented as a *PI-diagram* (Fig. 3A), which illustrates how the total information that $\mathbf{R}$ provides about S is distributed across various combinations of sources. Also, comparing this diagram with Fig. 1 makes immediately clear the meaning of each PI-term. First, from Theorem 4, we have that $I_\partial(S; \{1\}\{2\}) = I_\cap(S; \{1\}\{2\})$, corresponding to the redundancy for R_1 and R_2 . The unique information for R_1 is given by $I_\partial(S; \{1\}) = I(S; R_1) - I_\cap(S; \{1\}\{2\})$, which is the total information from

⁵The result that the Möbius inverse of a lattice function is nonnegative if, and only if, that function is totally monotone is proven in [22] and discussed further in [23]. Totally monotone functions are also called totally positive [22], [23] or strongly increasing [24] functions.

⁶Supermodularity means that Eq. (9) holds for $k = 2$.

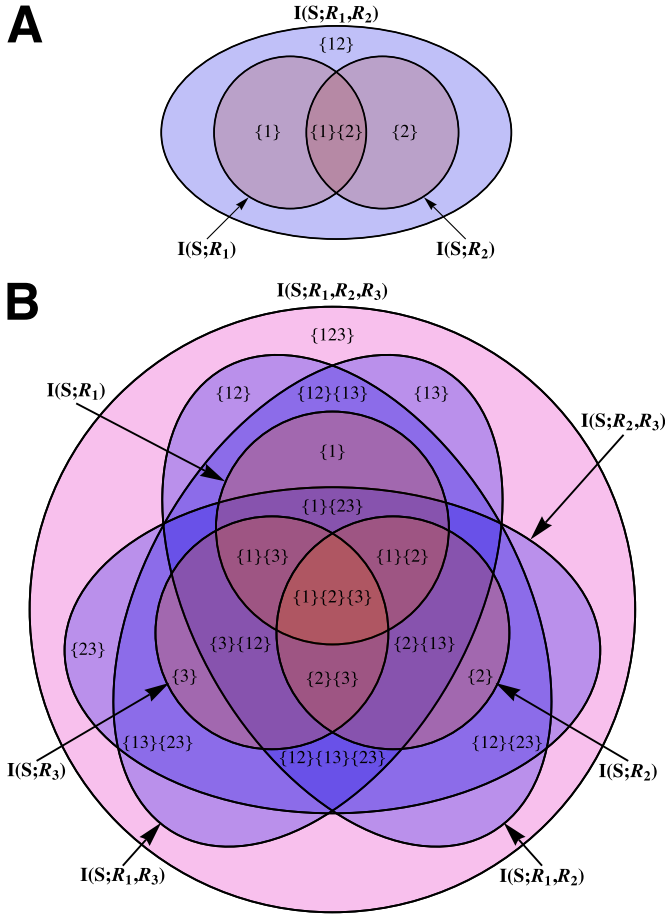


Fig. 3. Partial information diagrams for (A) 3 and (B) 4 variables.

R_1 minus the redundancy with R_2 , and likewise for R_2 . Finally, the additional information provided by the combination of R_1 and R_2 is given by $I_\partial(S; \{12\})$, corresponding to their synergy.

The general structure of PI-diagrams becomes clear when we consider the decomposition for four variables (Fig. 3B). First, note that all of the possibilities for three variables are again present for four. In particular, each element of $\mathbf{R}$ can provide unique information (regions labeled $\{1\}$, $\{2\}$, and $\{3\}$), information redundantly with one other variable ($\{1\}\{2\}$, $\{1\}\{3\}$, and $\{2\}\{3\}$), or information synergistically with one other variable ($\{12\}$, $\{13\}$, and $\{23\}$). Additionally, information can be provided redundantly by all three variables ($\{1\}\{2\}\{3\}$) or provided by their three-way synergy ($\{123\}$). More interesting are the new kinds of terms representing combinations of redundancy and synergy. For instance, the regions marked $\{1\}\{23\}$, $\{2\}\{13\}$, and $\{3\}\{12\}$ represent information that is available redundantly from either one variable considered individually or the other two considered together. Or, for instance, the region labeled $\{12\}\{13\}\{23\}$ represents the information provided redundantly by the three possible two-way synergies. In general, the PI-atom for a collection of sources corresponds to the information provided redundantly by the synergies of all sources in the collection. This point also clarifies our earlier claim that unique information is

best thought of as a degenerate case: unique information corresponds to the combination of first-order redundancy and first-order synergy.

In general, a PI-diagram for n variables, S and $\mathbf{R} = \{R_1, R_2, \dots, R_{n-1}\}$, consists of the following. First, for each element $R_i \in \mathbf{R}$ there is a region corresponding to $I(S; R_i)$. Then, for every subset $\mathbf{A}$ of $\mathbf{R}$ with two or more elements, $I(S; \mathbf{A})$ is depicted as a region containing $I(S; A)$ for all $A \in \mathbf{A}$ but not coextensive with $\bigcup_{A \in \mathbf{A}} I(S; A)$. The difference between $I(S; \mathbf{A})$ and $\bigcup_{A \in \mathbf{A}} I(S; A)$ represents the synergy for $\mathbf{A}$, the information gained from the combined knowledge of all elements in $\mathbf{A}$ that is not available from any subset. In addition, regions of the diagram intersect generically, representing all possibilities for redundancy. In total, a PI-diagram is composed of the $(n-1)$ -th Dedekind number [31] of PI-atoms, same as the cardinality of $\mathcal{A}(\mathbf{R})$ (see Appendix D). As described above, each PI-atom represents the redundancy of synergies for a particular collection of sources, corresponding to one distinct way for the components of $\mathbf{R}$ to contribute information about S .

PI-diagrams represent geometrically the underlying set-theoretic structure of multivariate information. In this context, $I_\cap$ is formally analogous to set intersection, with $I_\cap(S; \mathbf{A}_1, \dots, \mathbf{A}_k)$ corresponding to the region $\bigcap_i I(S; \mathbf{A}_i)$. This relationship also connects the redundancy lattice and PI-diagram for n variables: for $\alpha, \beta \in \mathcal{A}(\mathbf{R})$, α is lower than β in the redundancy lattice if, and only if, $\bigcap_{\mathbf{A} \in \alpha} I(S; \mathbf{A})$ is a subset of $\bigcap_{\mathbf{B} \in \beta} I(S; \mathbf{B})$ in the PI-diagram. If $I_\cap$ has a nonnegative PI-function (as is the case, for instance, with $I_{\min}$), then PI-diagrams represent a measure, in the formal sense of measure theory, over sets representing the information from each source. It is important to contrast the measure represented by PI-diagrams with that represented by I -diagrams, which were introduced by Yeung [15], [32] to formalize a set-theoretic view of Shannon information. In an I -diagram for n variables, sets represent the entropy for each variable and the intersection between sets represents the interaction information for those variables. The crucial difference between PI-diagrams and I -diagrams is that I -diagrams represent *signed* measures, meaning that the area of regions in an I -diagram can represent negative values. This is equivalent to the property that interaction information is sometimes negative, an idea that we revisit in the next section.

Now we turn from the general PI-function that is the inverse of $I_\cap$ to the specific PI-function that is the inverse of $I_{\min}$. We denote the PI-function corresponding to $I_{\min}$ by Π , so that

$$I_{\min}(S; \alpha) = \sum_{\beta \preceq \alpha} \Pi(S; \beta). \quad (13)$$

Theorem 5. Π is nonnegative.

As a first example, consider the distribution in Fig. 4A. From the symmetry of the distribution, it is clear that R_1 and R_2 must provide the same amount of information about S . Indeed, this is easily verified, with $I(S; R_1) = I(S; R_2) = -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3}$. However, it is also clear that R_1 and R_2 provide information about different outcomes of S . In particular, given knowledge of R_1 , one can determine conclusively

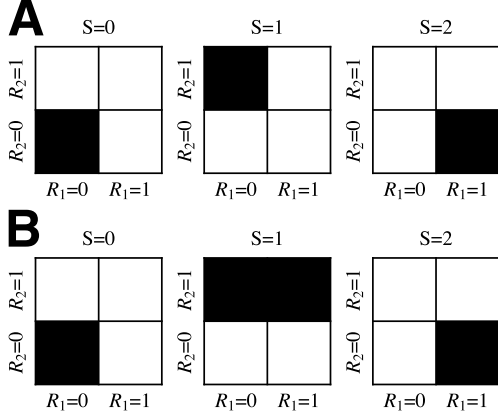


Fig. 4. Probability distributions for $S \in \{0, 1, 2\}$ and $R_1, R_2 \in \{0, 1\}$. Black tiles represent equiprobable outcomes. White tiles are zero-probability outcomes.

whether or not outcome $S = 2$ occurs (which is not the case for R_2), and likewise for R_2 and outcome $S = 1$. This is captured by $\Pi(S; \{1\}) = \Pi(S; \{2\}) = \frac{1}{3}$, meaning that R_1 and R_2 each provide $\frac{1}{3}$ bits of unique information about S . The redundant information, $\Pi(S; \{1\}\{2\}) = \log 3 - \log 2$, captures the fact that both R_1 or R_2 reduce uncertainty about S from three equally likely outcomes to two. Finally, R_1 and R_2 also provide $\frac{1}{3}$ bits of synergistic information, i.e., $\Pi(S; \{12\}) = \frac{1}{3}$. This value reflects the fact that R_1 and R_2 together uniquely determine whether or not outcome $S = 0$ occurs, which is not true for R_1 or R_2 alone.

Note that, unlike mutual information or interaction information, partial information is *not* symmetric. For instance, the synergistic information that R_1 and R_2 provide about S is not in general equal to the synergistic information that S and R_2 provide about R_1 . This property is also illustrated by the example in Fig. 4A. Given knowledge of S , one can uniquely determine the outcome of R_1 (and R_2), so that S provides complete information about both variables. Thus, it is not possible for the combination of S and R_2 to provide any additional synergistic information about R_1 , since there is no remaining uncertainty about R_1 when S is known. In contrast, as was just noted, R_1 and R_2 provide $\frac{1}{3}$ bits of synergistic information about S . This asymmetry accounts for our decision to focus on information *about* a particular variable S throughout, since in general the analysis will differ depending on the variable of interest. Note that total information is also asymmetric in this sense⁷, i.e., in general $I(S; R_1, R_2) \neq I(R_1; S, R_2)$.

Finally, we note two additional properties that are useful for the practical computation of PI-terms. First, the following theorem provides a simpler expression for Π :

Theorem 6.

$$\Pi(S; \alpha) = I_{\min}(S; \alpha) - \sum_s p(s) \max_{\beta \in \alpha^-} \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}).$$

Unlike Theorem 4, this expression for $\Pi(S; \alpha)$ only requires

⁷Though, of course, it is symmetric in the sense that $I(S; R_1, R_2) = I(R_1, R_2; S)$.

knowledge of the nodes *immediately* below α in the redundancy lattice (the set α^-). Second, in the same way that $I_{\min}$ acts like set intersection for PI-diagrams, $I_{\max}$ acts like set union, where $I_{\max}$ is defined exactly the same as $I_{\min}$ except substituting max for min. This is captured by the following theorem, which links $I_{\max}$ and $I_{\min}$ by an inclusion-exclusion relationship that is analogous to that between $\cup$ and $\cap$:

Theorem 7.

$$I_{\max}(S; \alpha) = \sum_{k=1}^{|\alpha|} (-1)^{k-1} \sum_{\substack{\beta \subseteq \alpha \\ |\beta|=k}} I_{\min}(S; \beta).$$

Thus, $I_{\max}(S; \mathbf{A}_1, \dots, \mathbf{A}_k)$ corresponds to the PI-diagram region $\bigcup_i I(S; \mathbf{A}_i)$, which has the appealing interpretation that the informational union of several sources equals the maximum information that is available from any one source. This correspondence between $I_{\min}/I_{\max}$ and $\cap/\cup$ means that expressions for PI-regions can be derived using the rich operations of set theory. For example, the synergy for all $(n-1)$ variables in $\mathbf{R}$ is conveniently given by

$$I(S; \mathbf{R}) - I_{\max}(S; \{\mathbf{A} \subset \mathbf{R} : |\mathbf{A}| = n-2\}). \quad (14)$$

Likewise, the unique information from a single R_i is given simply by

$$I(S; R_i) - I_{\min}(S; R_i, \mathbf{R} \setminus R_i). \quad (15)$$

V. WHY INTERACTION INFORMATION IS SOMETIMES NEGATIVE

We next show how PI-decomposition can be used to understand the conditions under which interaction information, the standard generalization of mutual information to multivariate interactions, is negative. The interaction information [3] for three variables is given by

$$I(S; R_1; R_2) = I(S; R_1|R_2) - I(S; R_1) \quad (16)$$

and for $n > 3$ variables is defined recursively as

$$I(S; R_1; R_2; \dots; R_{n-1}) = I(S; R_1; R_2; \dots; R_{n-2}|R_{n-1}) - I(S; R_1; R_2; \dots; R_{n-2}) \quad (17)$$

where the conditional interaction information is defined by simply including the conditioning in all terms of the original definition. Interaction information is symmetric for all permutations of its arguments, and is traditionally interpreted as the information shared by all n variables beyond that which is shared by any subset of those variables [7], [13], [15]–[18].

For 3-variable interaction information, a positive value is naturally interpreted as indicating a situation in which any one variable of the system enhances the correlation between the other two. For example, a positive value for Equation (16) indicates that knowledge of R_2 enhances the correlation between S and R_1 (and likewise for all other variable permutations). Thus, in the terminology used here, a positive value for $I(S; R_1; R_2)$ indicates the presence of synergy. On the other hand, a negative value for $I(S; R_1; R_2)$ indicates a situation in which any one variable accounts for or “explains away” [33] the correlation between the other two. In other

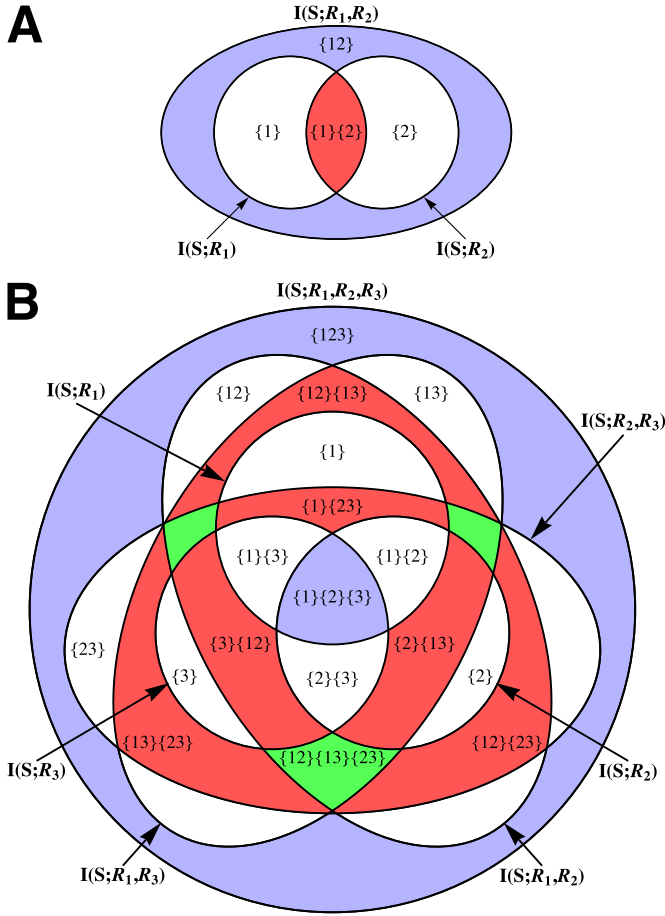


Fig. 5. PI-decomposition of interaction information for (A) 3 and (B) 4 variables. Blue and red regions represent PI-terms that are added and subtracted, respectively. The green region in (B) represents a PI-term that is subtracted twice.

words, a negative value for $I(S; R_1; R_2)$ indicates redundancy. Indeed, $I(S; R_1; R_2)$ is a widely used measure of synergy in neuroscience, where it is interpreted in exactly this way [34]–[37].

The PI-decomposition for 3-variable interaction information (Fig. 5A) confirms this interpretation. From Eqs. (11), (12) and (16), we have that $I(S; R_1; R_2)$ is equal to the difference between the synergistic and the redundant information, i.e.,

$$I(S; R_1; R_2) = \Pi(S; \{12\}) - \Pi(S; \{1\}\{2\}). \quad (18)$$

Thus, it is indeed the case that positive values indicate synergy and negative values indicate redundancy.

However, the PI-decomposition also makes clear that $I(S; R_1; R_2)$ confounds redundancy and synergy, with the meaning of interaction information ambiguous for any system that exhibits a mixture of the two (cf. [38], who suggest the possibility of mixed redundancy and synergy, but without attempting to disentangle them). For instance, consider again the example in Fig. 4A. As described earlier, R_1 and R_2 provide $\log 3 - \log 2$ bits of redundant information and $\frac{1}{3}$ bits of synergistic information. Thus, $I(S; R_1; R_2)$ is negative because there is more redundancy than synergy, despite the fact that the system clearly exhibits synergistic interactions. As a second example, consider the distribution in Fig. 4B. In

this case, R_1 and R_2 provide $\frac{1}{2}$ bits of redundant information, corresponding to the fact that R_1 or R_2 both reduce uncertainty about the outcomes $S = 0$ and $S = 2$. Additionally, R_1 and R_2 provide $\frac{1}{2}$ bits of synergistic information, reflecting the fact that R_1 and R_2 together provide complete information about outcomes $S = 0$ and $S = 2$, which is not true for either alone. Thus, the interaction information in this case is equal to zero despite the presence of both redundant and synergistic interactions, because redundancy and synergy are balanced.

The situation is worse for four-variable interaction information, which is known to violate the interpretation that positive values indicate (pure) synergy and negative values indicate (pure) redundancy [13], [39]. To demonstrate, consider the case of 3-parity, which is the higher-order form of the exclusive-OR, or 2-parity, function mentioned earlier. In this case, we have a system of four binary random variables, S and $\mathbf{R} = \{R_1, R_2, R_3\}$, where the eight outcomes for $\mathbf{R}$ are equiprobable and $S = R_1 \oplus R_2 \oplus R_3$. Intuitively, this corresponds to a case of pure synergy, since the value of S can be determined only when all of the R_i are known. Indeed, using Eq. (17) we find that $I(S; R_1; R_2; R_3)$ is equal to +1 bit, as expected from the interpretation that positive values indicate synergy. However, now consider a second system of binary variables, this time where the two outcomes of S are equiprobable and R_1, R_2 , and R_3 are all copies of S . Clearly this corresponds to pure redundancy, since the value of S can be determined uniquely from any R_i , but $I(S; R_1; R_2; R_3)$ for this system is again equal to +1 bit, same as the case of pure synergy. Thus, a completely redundant system is assigned a positive value for the interaction information, clearly violating the idea that redundancy is indicated by negative values. Worse still, the 4-variable interaction information fails to distinguish between the polar opposites of purely synergistic and purely redundant information.

The PI-decomposition for 4-variable interaction information (Fig. 5B) clarifies why this is the case. From Eqs. (10) and (17), it follows that $I(S; R_1; R_2; R_3)$ is equal to

$$\begin{aligned} & \Pi(S; \{123\}) + \Pi(S; \{1\}\{2\}\{3\}) \\ & - \Pi(S; \{1\}\{23\}) - \Pi(S; \{2\}\{13\}) - \Pi(S; \{3\}\{12\}) \\ & - \Pi(S; \{12\}\{13\}) - \Pi(S; \{12\}\{23\}) - \Pi(S; \{13\}\{23\}) \\ & - 2 \times \Pi(S; \{12\}\{13\}\{23\}). \end{aligned} \quad (19)$$

Thus, $I(S; R_1; R_2; R_3)$ equals the sum of third-order synergy ($\{123\}$) and third-order redundancy ($\{1\}\{2\}\{3\}$), minus the information provided redundantly by any first- and second-order synergy ($\{1\}\{23\}$, $\{2\}\{13\}$, and $\{3\}\{12\}$), minus the information provided redundantly by any two second-order synergies ($\{12\}\{13\}$, $\{12\}\{23\}$, and $\{13\}\{23\}$), and minus twice the information provided redundantly by all three second-order synergies ($\{12\}\{13\}\{23\}$). Thus, systems with pure synergy and pure redundancy have the same value for $I(S; R_1; R_2; R_3)$ because 4-variable interaction information adds in the highest-order synergy and redundancy terms. More generally, the PI-decomposition for $I(S; R_1; R_2; R_3)$ shows why it is difficult to interpret as a meaningful quantity, and as one might expect the story only becomes more complicated in higher dimensions. Thus, although one can readily

decompose interaction information into a collection of partial information contributions, and understand the conditions under which it will be positive or negative depending on the relative magnitudes of these contributions, the utility of interaction information for higher-dimensional systems is unclear.

VI. DISCUSSION

The main objective of this paper has been to quantify multivariate information in a way that illuminates the structure of variable interactions. This was accomplished by first defining an axiomatic measure of redundancy, $I_\cap$, that captures the intuition of redundancy as overlapping information. Next, it was shown that $I_\cap$ induces a lattice over the set of possible information sources, referred to as the redundancy lattice, which characterizes the distinct ways that information can be distributed amongst a set of sources. From this lattice, a measure of partial information was derived that captures the unique information contributed by each possible combination of sources. It was then shown that mutual information decomposes into a sum of these partial information terms, so that the total information provided by a source is broken down into a collection of partial information contributions. Moreover, it was demonstrated that each of these terms supports clear interpretation as a particular combination of redundant and synergistic interactions between specific subsets of variables. Finally, we discussed the relationship between PI-decomposition and interaction information, the current de facto measure of multivariate interactions, and used partial information to clarify the confusing property that interaction information is sometimes negative.

One obvious challenge with applying these ideas is that the number of PI-terms grows rapidly for larger systems. For instance, with 9 variables there are more than 5×10^{22} possibilities [40], and beyond that the Dedekind numbers are not even currently known. Thus, clearly an important direction for future work is to determine efficient ways of calculating partial information for larger systems. To this end, the lattice structure of the terms is likely to play an essential role. As with any ordered data structure, the fact that the space of possibilities is highly organized can be readily exploited for efficient use. For instance, as a simple example, if $I_{\min}$ is calculated in a descending fashion over the nodes of the redundancy lattice and at a certain juncture has a value of zero, all of the terms below that node can immediately be eliminated simply from the monotonicity of $I_{\min}$. Moreover, if the Markov property or any other constraints hold between the variables, many of the possible PI-terms can also be excluded. Finally, these considerations notwithstanding, it should also be emphasized that 3-variable interaction is the current state of the art, and thus even the simplest form of PI-decomposition can be used to address a number of outstanding questions.

In physics, for example, 3-variable interactions have been explored in relation to the non-separability of quantum systems [41] and in the study of many-body correlation effects [42]. In neuroscience, the concepts of synergy and redundancy for three variables have been examined in the context of neural coding in a number of theoretical and empirical investigations

[34]–[37], [43], [44]. In genetics, multivariate dependencies arise in the analysis of gene-gene and gene-environment interactions in studies of human disease susceptibility [39], [45], [46]. Similar issues have also been explored in machine learning [33], [38], [47], ecology [48], quantum information theory [49], information geometry [9], rough set analysis [50], and cooperative game theory [51]. In addition, the specific application of PI-decomposition to multivariate time series analysis, providing a general method for quantifying information transfer in nonlinear dynamical systems, has been described recently in [52]. In all of these cases, the 3-variable form of PI-decomposition can be applied immediately to illuminate the structure of multivariate dependencies, while the general form provides a clear way forward in the study of more complex systems of interactions.

APPENDIX A MEASURES OF SPECIFIC INFORMATION

Measures of specific information are discussed in [53] in the context of quantifying the information that specific neural responses provide about a stimulus ensemble. For random variables S and R , representing stimuli and responses, respectively, the information that R provides about S is decomposed according to

$$I(S; R) = \sum_{r \in R} p(r) i_r(r) \quad (20)$$

and

$$i_r(r) = H(S) - H(S|r) \quad (21)$$

where $H(S)$ is the entropy of S and $i_r(r)$ is the *response-specific information* associated with each $r \in R$. The response-specific information quantifies the change in uncertainty about S when response r is observed. In [53], it is shown that i_r is the only measure of specific information that is additive, though it is also possible for i_r to be negative.

To distinguish the different role played by stimuli as opposed to responses, an alternative measure of specific information is proposed in [54]. The *stimulus-specific information* for an outcome $s \in S$ is defined as

$$i_s(s) = \sum_{r \in R} p(r|s) i_r(r). \quad (22)$$

Like the response-specific information, the weighted average of $i_s(s)$ gives the mutual information $I(S; R)$. Stimulus-specific information quantifies the extent to which a particular stimulus s tends to evoke responses that are informative about the entire ensemble S (responses with high values for i_r).

Finally, both [53] and [54] also discuss $I(S = s; R)$, the measure of specific information used here (Eq. (3)). In [54], $I(S = s; R)$ is described as the reduction in surprise of a particular stimulus s gained from each response, averaged over all responses associated with that stimulus. Thus, whereas $i_s(s)$ weights each response r according to the information that it contributes about the entire ensemble S , $I(S = s; R)$ quantifies only the information that R provides about the particular outcome $S = s$. In [53], it is proven that $I(S = s; R)$ is the only measure of specific information that is nonnegative.

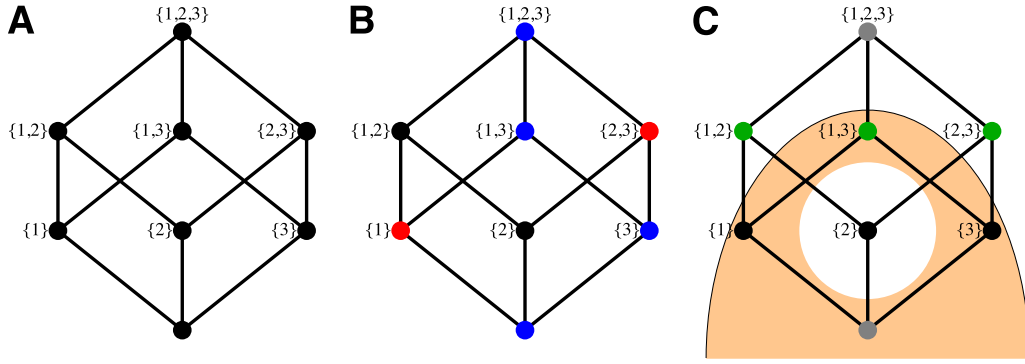


Fig. 6. Basic lattice-theoretic concepts. (A) Hasse diagram of the lattice $\langle \mathcal{P}(X), \subseteq \rangle$ for $X = \{1, 2, 3\}$. (B) An example of a chain (blue nodes) and an antichain (red nodes). (C) The top $\top$ and bottom $\perp$ are shown in gray. Green nodes correspond to $\{1, 2, 3\}^-$, the set of elements covered by $\{1, 2, 3\}$. The orange region represents $\downarrow \{1, 3\}$, the down-set of $\{1, 3\}$.

APPENDIX B REDUNDANCY AND DEPENDENCE

If sources provide redundant information, must they themselves be dependent? Conversely, if sources are dependent, do they necessarily provide redundant information? It turns out that the answer to both questions is no, as the following simple examples illustrate. Let R_1 and R_2 be binary and uniformly distributed. If $R_1 = R_2$ and S is a degenerate random variable, then $I(R_1; R_2) > 0$ and $I_{\min}(S; R_1, R_2) = 0$, so that dependence does not imply redundancy. On the other hand, if R_1 and R_2 are independent and $S = R_1 \wedge R_2$, then $I_{\min}(S; R_1, R_2) > 0$ and $I(R_1; R_2) = 0$, so that redundancy does not imply dependence. Considering instead conditional dependence, if R_1 and R_2 are independent and $S = R_1 \oplus R_2$, then $I(R_1; R_2 | S) > 0$ and $I_{\min}(S; R_1, R_2) = 0$, so that conditional dependence does not imply redundancy. Conversely, if $S = R_1 = R_2$, then $I_{\min}(S; R_1, R_2) > 0$ and $I(R_1; R_2 | S) = 0$, so that redundancy does not imply conditional dependence.

APPENDIX C LATTICE THEORY DEFINITIONS

Here we review only the basic concepts of lattice theory needed for the supporting proofs. For a thorough treatment, see [55], [56].

Definition 1. A pair $\langle X, \leq \rangle$ is a partially ordered set or poset if $\leq$ is a binary relation on X that is reflexive, transitive and antisymmetric.

Definition 2. Let $Y \subseteq X$. Then $a \in Y$ is a maximal element in Y if for all $b \in Y, a \leq b \Rightarrow a = b$. A minimal element is defined dually. We denote the set of maximal elements of Y by $\overline{Y}$ and the set of minimal elements by $\underline{Y}$.

Definition 3. Let $\langle X, \leq \rangle$ be a poset, and let $Y \subseteq X$. An element $x \in X$ is an upper bound for Y if for all $y \in Y, y \leq x$. A lower bound for Y is defined dually.

Definition 4. An element $x \in X$ is the least upper bound or supremum for Y , denoted $\sup Y$, if x is an upper bound of Y and for all $y \in Y$ and all $z \in X, y \leq z$ implies $x \leq z$.

The greatest upper bound or infimum for Y , denoted $\inf Y$, is defined dually.

Definition 5. A poset $\langle X, \leq \rangle$ is a lattice if, and only if, for all $x, y \in X$ both $\inf\{x, y\}$ and $\sup\{x, y\}$ exist in X . If $\langle X, \leq \rangle$ is a lattice, it is common to write $x \wedge y$, the meet of x and y , and $x \vee y$, the join of x and y , for $\inf\{x, y\}$ and $\sup\{x, y\}$, respectively. For $Y \subseteq X$, we use $\bigwedge Y$ and $\bigvee Y$ to denote the meet and join of all elements in Y , respectively.

Definition 6. For $a, b \in X$, we say that a is covered by b (or b covers a) if $a < b$ and $a \leq c < b \Rightarrow a = c$. The set of elements that are covered by b is denoted by b^- .

The classic example of a lattice is the power set of a set X ordered by inclusion, denoted $\langle \mathcal{P}(X), \subseteq \rangle$. Lattices are naturally represented by Hasse diagrams, in which nodes correspond to members of X and an edge exists between elements x and y if x covers y . Fig. 6A depicts the Hasse diagram for the lattice $\langle \mathcal{P}(X), \subseteq \rangle$ with $X = \{1, 2, 3\}$.

Definition 7. If $\langle X, \leq \rangle$ is a poset, $Y \subseteq X$ is a chain if for all $a, b \in Y$ either $a \leq b$ or $b \leq a$. Y is an antichain if $a \leq b$ only if $a = b$.

Fig. 6B shows examples of a chain and an antichain.

Definition 8. If there exists an element $\perp \in X$ with the property that $\perp \leq x$ for all $x \in X$, we call $\perp$ the bottom element of X . The top element of X , denoted by $\top$, is defined dually.

Definition 9. For any $x \in X$, we define

$$\downarrow x = \{y \in X : y \leq x\} \text{ and } \dot{\downarrow} x = \{y \in X : y < x\}$$

where $\downarrow x$ and $\dot{\downarrow} x$ are called the down-set and strict down-set of x , respectively.

Fig. 6C illustrates the concepts of top and bottom elements, covering relations, and down-sets.

APPENDIX D $\mathcal{A}(\mathbf{R})$ AND THE REDUNDANCY LATTICE

As mentioned in the main text, $\mathcal{A}(\mathbf{R})$ corresponds to the set of antichains on the lattice $\langle \mathcal{P}(\mathbf{R}), \subseteq \rangle$, excluding the empty

set. The cardinality of this set for $|\mathbf{R}| = n - 1$ is given by the $(n - 1)$ -th Dedekind number, which for $n = 2, 3, 4, \dots$ is $1, 4, 18, 166, 7579, \dots$ ([31], p. 273). The fact that $\langle \mathcal{A}(\mathbf{R}), \preceq \rangle$ forms a lattice, which we call the redundancy lattice, is proven in [57], where the corresponding lattice is denoted $\langle \mathcal{A}(X), \preceq' \rangle$ (see also [58]). As shown in [57], the meet ($\wedge$) and join ($\vee$) for this lattice are given by

$$\alpha \wedge \beta = \underline{\alpha \cup \beta} \quad (23)$$

and

$$\alpha \vee \beta = \underline{\uparrow \alpha \cap \uparrow \beta}. \quad (24)$$

APPENDIX E SUPPORTING PROOFS

Lemma 1. $I(S = s; A) \geq 0$ with equality if, and only if,

$$p(s, a) = p(s)p(a) \quad \forall a \in A.$$

Proof:

$$I(S = s; A) = D(p(a|s) \parallel p(a)) \geq 0$$

where $D(\cdot \parallel \cdot)$ is the Kullback-Leibler divergence and the second step follows from the information inequality ([2], p. 26). $D(q \parallel r) = 0$ if, and only if, $q = r$, so $I(S = s; A) = 0$ if, and only if,

$$p(s, a) = p(s)p(a) \quad \forall a \in A.$$

□

Lemma 2. If $\mathbf{A} \subseteq \mathbf{B}$, then $I(S = s; \mathbf{A}) \leq I(S = s; \mathbf{B})$.

Proof: Consider $\mathbf{A} \subset \mathbf{B} \subseteq \mathbf{R}$ and let $\mathbf{C} = \mathbf{B} \setminus \mathbf{A}$. Then we have

$$\begin{aligned} & I(S = s; \mathbf{B}) - I(S = s; \mathbf{A}) \\ &= \sum_{\mathbf{b}} p(\mathbf{b}|s) \log \frac{p(s, \mathbf{b})}{p(s)p(\mathbf{b})} - \sum_{\mathbf{a}} p(\mathbf{a}|s) \log \frac{p(s, \mathbf{a})}{p(s)p(\mathbf{a})} \\ &= \sum_{\mathbf{a}} \sum_{\mathbf{c}} p(\mathbf{a}, \mathbf{c}|s) \log \frac{p(s, \mathbf{a}, \mathbf{c})}{p(s)p(\mathbf{a}, \mathbf{c})} - \\ & \quad \sum_{\mathbf{a}} \sum_{\mathbf{c}} p(\mathbf{a}, \mathbf{c}|s) \log \frac{p(s, \mathbf{a})}{p(s)p(\mathbf{a})} \\ &= \sum_{\mathbf{a}} \sum_{\mathbf{c}} p(\mathbf{a}, \mathbf{c}|s) \log \frac{p(s, \mathbf{c}|\mathbf{a})}{p(s|\mathbf{a})p(\mathbf{c}|\mathbf{a})} \\ &= \sum_{\mathbf{a}} p(\mathbf{a}) \sum_{\mathbf{c}} p(\mathbf{c}|\mathbf{a}, s) \log \frac{p(\mathbf{c}|\mathbf{a}, s)}{p(\mathbf{c}|\mathbf{a})} \\ &= \sum_{\mathbf{a}} p(\mathbf{a}) D(p(\mathbf{c}|\mathbf{a}, s) \parallel p(\mathbf{c}|\mathbf{a})) \geq 0. \end{aligned}$$

□

Theorem 1. $I_{\min}$ satisfies the axiomatic definition $I_{\cap}$.

Proof: That $I_{\min}$ satisfies Axioms 1 and 2 follows from basic properties of min, while Axiom 3 follows from Lemma 2. □

Theorem 2. $I_{\min}(S; \mathbf{A}_1, \dots, \mathbf{A}_k) = 0$ if, and only if, for every joint outcome $(s, \mathbf{a}_1, \dots, \mathbf{a}_k)$,

$$p(s, \mathbf{a}_i) = p(s)p(\mathbf{a}_i)$$

for at least one $\mathbf{a}_i$.

Proof: It follows trivially from Lemma 1 and basic properties of min. □

Theorem 3. If $S \rightarrow \mathbf{A}_k \rightarrow \mathbf{A}_{k-1}$, then

$$I_{\min}(S; \mathbf{A}_1, \dots, \mathbf{A}_{k-1}, \mathbf{A}_k) = I_{\min}(S; \mathbf{A}_1, \dots, \mathbf{A}_{k-1}).$$

Proof: It suffices to show that if $S \rightarrow \mathbf{A}_k \rightarrow \mathbf{A}_{k-1}$, then $I(S = s; \mathbf{A}_{k-1}) \leq I(S = s; \mathbf{A}_k)$ for all $s \in S$. If $S \rightarrow \mathbf{A}_k \rightarrow \mathbf{A}_{k-1}$, then $I(S; \mathbf{A}_{k-1}|\mathbf{A}_k) = 0$ and thus $I(S; \mathbf{A}_{k-1}, \mathbf{A}_k) = I(S; \mathbf{A}_k)$. From Lemma 2, this means that $I(S = s; \mathbf{A}_{k-1}, \mathbf{A}_k) = I(S = s; \mathbf{A}_k)$ for all $s \in S$. Also from Lemma 2, we have that

$$I(S = s; \mathbf{A}_{k-1}, \mathbf{A}_k) \geq I(S = s; \mathbf{A}_{k-1}),$$

so we conclude that $I(S = s; \mathbf{A}_{k-1}) \leq I(S = s; \mathbf{A}_k)$ for all $s \in S$. □

Theorem 4.

$$I_{\partial}(S; \alpha) = I_{\cap}(S; \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathcal{B} \subseteq \alpha^- \\ |\mathcal{B}|=k}} I_{\cap}(S; \bigwedge \mathcal{B})$$

where $\bigwedge$ is the meet operator and α^- represents the nodes immediately below α on the redundancy lattice.

Proof: For $\mathcal{B} \subseteq \mathcal{A}(\mathbf{R})$, define the set-additive function f as

$$f(\mathcal{B}) = \sum_{\beta \in \mathcal{B}} I_{\partial}(S; \beta).$$

From Eq. (7), it follows that $I_{\cap}(S; \alpha) = f(\downarrow \alpha)$ and

$$\begin{aligned} I_{\partial}(S; \alpha) &= f(\downarrow \alpha) - f(\downarrow \alpha) \\ &= f(\downarrow \alpha) - f\left(\bigcup_{\beta \in \alpha^-} \downarrow \beta\right). \end{aligned}$$

Applying the inclusion-exclusion principle ([21], p. 64), we have

$$= f(\downarrow \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathcal{B} \subseteq \alpha^- \\ |\mathcal{B}|=k}} f\left(\bigcap_{\gamma \in \mathcal{B}} \downarrow \gamma\right)$$

and it is a basic result of lattice theory that for any lattice L and $A \subseteq L$, $\bigcap_{a \in A} \downarrow a = \downarrow (\bigwedge A)$ ([55], p. 57), so we have

$$\begin{aligned} &= f(\downarrow \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathcal{B} \subseteq \alpha^- \\ |\mathcal{B}|=k}} f(\downarrow (\bigwedge \mathcal{B})) \\ &= I_{\cap}(S; \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\mathcal{B} \subseteq \alpha^- \\ |\mathcal{B}|=k}} I_{\cap}(S; \bigwedge \mathcal{B}). \end{aligned}$$

□ Then, applying Lemma 3 we have

$$= I_{\min}(S; \alpha) - \sum_s p(s) \max_{\beta \in \alpha^-} \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}).$$

□

Theorem 5. Π is nonnegative.

Proof: If $\alpha = \perp$, $\Pi(S; \alpha) = I_{\min}(S; \alpha)$ and $\Pi(S; \alpha) \geq 0$ follows from the nonnegativity of $I_{\min}$. To prove it for $\alpha \neq \perp$, we proceed by contradiction. Assume there exists $\alpha \in \mathcal{A}(\mathbf{R}) \setminus \{\perp\}$ such that $\Pi(S; \alpha) < 0$. Beginning with the simpler expression for $\Pi(S; \alpha)$ proven in Theorem 6 and combining summations yields

$$\Pi(S; \alpha) = \sum_s p(s) \left\{ \min_{\mathbf{A} \in \alpha} I(S = s; \mathbf{A}) - \max_{\beta \in \alpha^-} \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}) \right\}.$$

From this equation, it is clear that there must exist $\beta \in \alpha^-$ such that for all $\mathbf{B} \in \beta$, $I(S = s; \mathbf{A}) < I(S = s; \mathbf{B})$ for some outcome $s \in S$ and some $\mathbf{A} \in \alpha$. Thus, from Lemma 2, there does not exist $\mathbf{B} \in \beta$ such that $\mathbf{B} \subseteq \mathbf{A}$. However, since $\beta \prec \alpha$ by definition, there exists $\mathbf{B} \in \beta$ such that $\mathbf{B} \subseteq \mathbf{A}$. □

Lemma 3 (Maximum-minimums identity). *Let A be a set of numbers. The maximum-minimums identity states that*

$$\max A = \sum_{k=1}^{|A|} (-1)^{k-1} \sum_{\substack{B \subseteq A \\ |B|=k}} \min B,$$

or conversely,

$$\min A = \sum_{k=1}^{|A|} (-1)^{k-1} \sum_{\substack{B \subseteq A \\ |B|=k}} \max B.$$

Proof: It is proven in a number of introductory texts, e.g. [59]. □

Theorem 6.

$$\Pi(S; \alpha) = I_{\min}(S; \alpha) - \sum_s p(s) \max_{\beta \in \alpha^-} \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}).$$

Proof:

Rewriting Theorem 4, $\Pi(S; \alpha)$ can be written as

$$\begin{aligned} & I_{\min}(S; \alpha) - \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\beta \in \alpha^- \\ |\beta|=k}} \sum_s p(s) \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}) \\ &= I_{\min}(S; \alpha) - \\ & \sum_s p(s) \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\beta \in \alpha^- \\ |\beta|=k}} \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}) \end{aligned}$$

and by Lemma 2 and Eq. (23),

$$\begin{aligned} &= I_{\min}(S; \alpha) - \\ & \sum_s p(s) \sum_{k=1}^{|\alpha^-|} (-1)^{k-1} \sum_{\substack{\beta \in \alpha^- \\ |\beta|=k}} \min_{\mathbf{B} \in \beta} \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}). \end{aligned}$$

Theorem 7.

$$I_{\max}(S; \alpha) = \sum_{k=1}^{|\alpha|} (-1)^{k-1} \sum_{\substack{\beta \subseteq \alpha \\ |\beta|=k}} I_{\min}(S; \beta).$$

Proof: Using Lemma 3, $I_{\max}$ can be written as

$$I_{\max}(S; \alpha) = \sum_s p(s) \sum_{k=1}^{|\alpha|} (-1)^{k-1} \sum_{\substack{\beta \subseteq \alpha \\ |\beta|=k}} \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}),$$

and rearranging the order of summations,

$$\begin{aligned} &= \sum_{k=1}^{|\alpha|} (-1)^{k-1} \sum_{\substack{\beta \subseteq \alpha \\ |\beta|=k}} \sum_s p(s) \min_{\mathbf{B} \in \beta} I(S = s; \mathbf{B}) \\ &= \sum_{k=1}^{|\alpha|} (-1)^{k-1} \sum_{\substack{\beta \subseteq \alpha \\ |\beta|=k}} I_{\min}(S; \beta). \end{aligned}$$

□

ACKNOWLEDGEMENT

The authors wish to thank Prof. J. Beggs, V. Griffith, A. Kolchinsky, Prof. O. Sporns, and Prof. L. Yaeger for their helpful comments.

REFERENCES

- [1] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Univ of Illinois Press, 1949.
- [2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.
- [3] W. J. McGill, "Multivariate information transmission," *Psychometrika*, vol. 19, no. 2, pp. 97–116, 1954.
- [4] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of Research and Development*, vol. 4, no. 1, pp. 66–82, 1960.
- [5] I. J. Good, "Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables," *Annals of Mathematical Statistics*, vol. 34, no. 3, pp. 911–934, 1963.
- [6] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Annals of Probability*, vol. 3, no. 1, pp. 146–158, 1975.
- [7] T. S. Han, "Multiple mutual informations and multiple interactions in frequency data," *Information and Control*, vol. 46, pp. 26–45, 1980.
- [8] M. Studený and J. Vejnarová, "The multiinformation function as a tool for measuring stochastic dependence," *Learning in Graphical Models*, pp. 261–297, 1998.
- [9] S. Amari, "Information geometry on hierarchical decomposition of stochastic interactions," *IEEE Transactions on Information Theory*, vol. 47, pp. 1701–1711, 2001.
- [10] E. Schneidman, S. Still, M. J. Berry, and W. Bialek, "Network information and connected correlations," *Physical Review Letters*, vol. 91, no. 23, p. 238701, 2003.
- [11] W. R. Garner, *Uncertainty and Structure as Psychological Concepts*. Wiley, 1962.

- [12] G. Tononi, O. Sporns, and G. M. Edelman, "A measure for brain complexity: relating functional segregation and integration in the nervous system," *Proceedings of the National Academy of Science USA*, vol. 91, no. 11, pp. 5033–5037, 1994.
- [13] A. J. Bell, "The co-information lattice," *Proceedings of ICA2003*, pp. 921–926, 2003.
- [14] T. J. Gawne and B. J. Richmond, "How independent are the messages carried by adjacent inferior temporal cortical neurons?" *Journal of Neuroscience*, vol. 13, no. 7, pp. 2758–2771, 1993.
- [15] R. W. Yeung, *Information Theory and Network Coding*. Springer, 2008.
- [16] S. Takano, "On 3-dimensional interaction information," *Proceedings of the Japan Academy*, vol. 50, no. 2, pp. 109–113, 1974.
- [17] T. Tsujishita, "On triple mutual information," *Advances in Applied Mathematics*, vol. 16, no. 3, pp. 269–274, 1995.
- [18] Z. Zhang and R. W. Yeung, "On characterization of entropy functions via information inequalities," *IEEE Transactions on Information Theory*, vol. 44, pp. 1440–1452, 1998.
- [19] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code*. MIT Press, 1999.
- [20] G. Rota, "On the foundations of combinatorial theory I. Theory of Möbius functions," *Probability Theory and Related Fields*, vol. 2, no. 4, pp. 340–368, 1964.
- [21] R. P. Stanley, *Enumerative Combinatorics*. Cambridge Univ Press, 1997, vol. 1.
- [22] I. Gilboa and E. Lehrer, "Global games," *International Journal of Game Theory*, vol. 20, no. 2, pp. 129–147, 1991.
- [23] G. Rossi, "Information functions and expectation," *Risk, Uncertainty and Decision*, pp. 1–30, 2004.
- [24] L. Lovász and A. Schrijver, "Cones of matrices and set-functions and 0-1 optimization," *SIAM Journal of Optimization*, vol. 1, pp. 166–190, 1991.
- [25] G. J. Klir, *Uncertainty and Information*. Wiley Online Library, 2006.
- [26] M. Grabisch, "Belief functions on lattices," *International Journal of Intelligent Systems*, vol. 24, no. 1, pp. 76–95, 2009.
- [27] F. Matúš, "Probabilistic conditional independence structures and matroid theory," *International Journal of General Systems*, vol. 22, no. 2, pp. 185–196, 1993.
- [28] M. Madiman and P. Tetali, "Information inequalities for joint distributions, with interpretations and applications," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2699–2713, 2010.
- [29] B. Steudel, D. Janzing, and B. Schölkopf, "Causal Markov condition for submodular information measures," *Arxiv preprint arXiv:1002.4020*, 2010.
- [30] B. Steudel and N. Ay, "Information-theoretic inference of common ancestors," *Arxiv preprint arXiv:1010.5720*, 2010.
- [31] L. Comtet, *Advanced Combinatorics: The Art of Finite and Infinite Expansions*. Springer, 1974.
- [32] R. W. Yeung, "A new outlook on Shannon's information measures," *IEEE Transactions on Information Theory*, vol. 37, no. 3, pp. 466–474, 1991.
- [33] J. Pearl and G. Shafer, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
- [34] N. Brenner, S. P. Strong, R. Koberle, W. Bialek, and R. de Ruyter van Steveninck, "Synergy in a neural code," *Neural Computation*, vol. 12, no. 7, pp. 1531–1552, 2000.
- [35] S. Panzeri, S. R. Schultz, A. Treves, and E. T. Rolls, "Correlations and the encoding of information in the nervous system," *Proceedings of the Royal Society B*, vol. 266, no. 1423, pp. 1001–1012, 1999.
- [36] E. Schneidman, W. Bialek, and M. J. Berry, "Synergy, redundancy, and independence in population codes," *Journal of Neuroscience*, vol. 23, no. 37, pp. 11 539–11 553, 2003.
- [37] P. E. Latham and S. Nirenberg, "Synergy, redundancy, and independence in population codes, revisited," *Journal of Neuroscience*, vol. 25, no. 21, pp. 5195–5206, 2005.
- [38] A. Jakulin and I. Bratko, "Quantifying and visualizing attribute interactions," *Arxiv preprint cs/0308002*, 2003.
- [39] D. Anastassiou, "Computational analysis of the synergy among multiple interacting genes," *Molecular Systems Biology*, vol. 3, no. 83, pp. 1–8, 2007.
- [40] D. Wiedemann, "A computation of the eighth Dedekind number," *Order*, vol. 8, no. 1, pp. 5–6, 1991.
- [41] N. J. Cerf and C. Adami, "Entropic Bell inequalities," *Physical Review A*, vol. 55, no. 5, pp. 3371–3374, 1997.
- [42] H. Matsuda, "Physical nature of higher-order mutual information: Intrinsic correlations and frustration," *Physical Review E*, vol. 62, no. 3, pp. 3096–3102, 2000.
- [43] I. Gat and N. Tishby, "Synergy and redundancy among brain cells of behaving monkeys," *Advances in NIPS*, pp. 111–117, 1999.
- [44] N. S. Narayanan, E. Y. Kimchi, and M. Laubach, "Redundancy and synergy of neuronal ensembles in motor cortex," *Journal of Neuroscience*, vol. 25, no. 17, pp. 4207–4216, 2005.
- [45] J. H. Moore, J. C. Gilbert, C. T. Tsai, F. T. Chiang, T. Holden, N. Barney, and B. C. White, "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility," *Journal of Theoretical Biology*, vol. 241, no. 2, pp. 252–261, 2006.
- [46] P. Chanda, A. Zhang, D. Brazeau, L. Sucheston, J. L. Freudenheim, C. Ambrosone, and M. Ramanathan, "Information-theoretic metrics for visualizing gene-environment interactions," *American Journal of Human Genetics*, vol. 81, no. 5, pp. 939–963, 2007.
- [47] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge Univ Press, 2003.
- [48] L. Orlóci, M. Anand, and V. D. Pillar, "Biodiversity analysis: issues, concepts, techniques," *Community Ecology*, vol. 3, no. 2, pp. 217–236, 2002.
- [49] V. Vedral, "The role of relative entropy in quantum information theory," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 197–234, 2002.
- [50] G. Gediga and I. Düntsch, "On model evaluation, indices of importance, and interaction values in rough set analysis," in *Rough-Neural Computing*, S. K. Pal, L. Polkowski, and A. Skowron, Eds. Heidelberg: Physica Verlag, 2003.
- [51] M. Grabisch and M. Roubens, "An axiomatic approach to the concept of interaction among players in cooperative games," *International Journal of Game Theory*, vol. 28, no. 4, pp. 547–565, 1999.
- [52] P. L. Williams and R. D. Beer, "Generalized measures of information transfer," *Arxiv preprint arXiv:1102.1507*, 2011.
- [53] M. R. DeWeese and M. Meister, "How to measure the information gained from one symbol," *Network*, vol. 10, pp. 325–340, 1999.
- [54] D. A. Butts, "How much information is associated with a particular stimulus?" *Network*, vol. 14, pp. 177–187, 2003.
- [55] B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order*, 2nd ed. Cambridge Univ Press, 2002.
- [56] G. A. Grätzer, *General Lattice Theory*, 2nd ed. Birkhäuser, 2003.
- [57] J. Crampton and G. Loizou, "Two partial orders on the set of antichains," September 2000, research note.
- [58] —, "The completion of a poset in a lattice of antichains," *International Mathematical Journal*, vol. 1, no. 3, pp. 223–238, 2001.
- [59] S. M. Ross, *A First Course in Probability*, 8th ed. Prentice Hall, 2009.