Supporting Information for

Partial entropy decomposition reveals higher-order structures in human brain activity

Thomas F. Varley, Maria Pope, Maria Grazia Puxeddu, Josh Faskowitz, Olaf Sporns

Corresponding Thomas F. Varley.
E-mail: tvarley@iu.edu

This PDF file includes:

1. Supplementary Figures

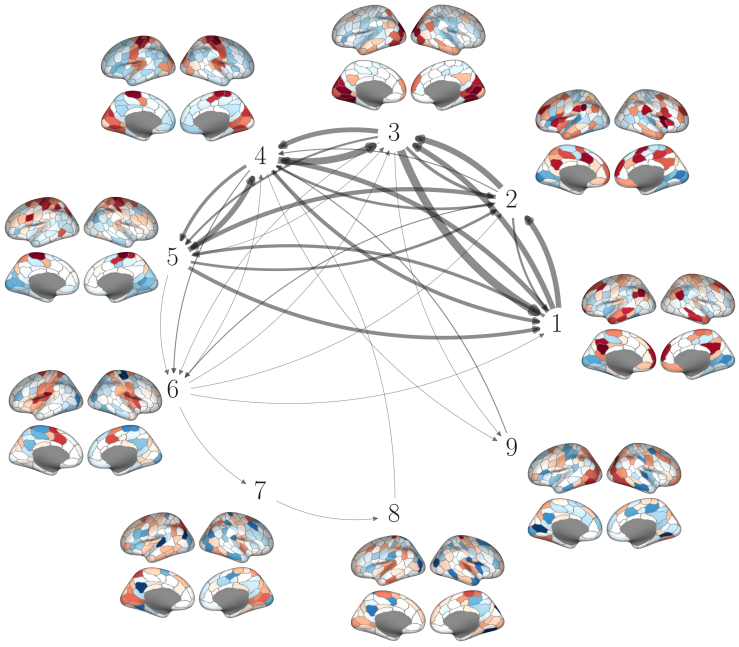

Fig. S1. State-to-state transitions. For each of the nine distinct states, we can see how many times each state transitions another (self-loops are not shown for visual clarity). We can see that the various states have meaningful differences between each-other (e.g. the visual system or the somato-motor systems both transition from redundancy- to synergy-dominated configurations over time), however, within a state, the patterns are largely symmetrical across hemispheres.

2. Partial Entropy Decomposition

The *partial entropy decomposition* (PED) provides a framework with which we can extract *all* of the meaningful structure in a system of interacting random variables (1). By structure, we are referring to the (possibly higher-order) patterns of information-sharing between elements. Consider a system $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$, comprised of N interacting, discrete random variables: the set of all informative relationships between elements (and ensembles of elements) in $\mathbf{X}$ forms its structure. We begin by defining the total entropy of $\mathbf{X}$ using the Shannon entropy:

$$\mathcal{H}(\mathbf{X}) := -\sum_{\mathbf{x} \in \mathfrak{X}} \mathcal{P}(\mathbf{x}) \log_2 \mathcal{P}(\mathbf{x}) \quad [1]$$

Where $\mathbf{x}$ indicates a particular configuration of $\mathbf{X}$ and $\mathfrak{X}$ is the support set of $\mathbf{X}$. This joint entropy quantifies, on average, how much it is possible to know about $\mathbf{X}$ (i.e. how many bits of information would be required, on average, to reduce our uncertainty to zero). The entropy is a summary statistic describing an entire distribution $\mathcal{P}(X)$:

$$\mathcal{H}(\mathbf{X}) = \mathbb{E}[-\log_2 \mathcal{P}(\mathbf{x})] \quad [2]$$

Where $-\log_2 \mathcal{P}(x)$ is the *local entropy* $h(\mathbf{x})$. We can intuitively understand the local entropy with the logic of local probability mass exclusions (2, 5). Suppose that we observe $\mathbf{X} = \mathbf{x}$. Upon observing $\mathbf{x}$, we can immediately *rule out* the possibility that $\mathbf{X}$ is in any state $\neg\mathbf{x}$, and by ruling out those possibilities, we exclude all the probability mass associated with $\mathcal{P}(\mathbf{X} = \neg\mathbf{x})$. If $\mathcal{P}(\mathbf{x})$ is very low, then upon learning $\mathbf{X} = \mathbf{x}$, we exclude a large amount of probability mass $(1 - \mathcal{P}(\mathbf{x}))$, and consequently, $h(\mathbf{x})$ is high. Conversely, if $\mathcal{P}(\mathbf{x})$ is large, then only a small amount of probability mass is excluded, and so $h(\mathbf{x})$ is low.

A. Quantifying Shared Entropy. The measure $h(\mathbf{x})$ is a very crude one: it gives us a single summary statistic that describes the behaviour of the whole without making reference to the structure of the relationships between $\mathbf{x}$'s constituent elements. If $\mathbf{X}$ has some non-trivial structure that integrates multiple elements (or ensembles of elements), then we propose that those elements must share entropy. This notion of shared entropy forms the cornerstone of the PED. The way all of the parts of $\mathbf{X}$ share entropy forms the structure of the system. In the original proposal of the PED by Ince (1), shared entropy ($\mathcal{H}_{cs}$) was defined using the local co-information, which treats the entropy of variables as sets and defines the shared entropy using inclusion-exclusion criteria. Unfortunately, as discussed by Finn and Lizier, the set-theoretic interpretation of mutlivariate mutual information is complex, as both the expected and local co-information can be negative (6), and the PED computed using Ince's proposed method can result in negative values that are difficult to interpret.

Here, we propose an alternative way to operationalize the notion of redundant entropy by saying that two variables $X_1, X_2 \in \mathbf{X}$ share entropy if they induce the same exclusions: i.e. if learning X_1 or X_2 rules out the same configurations of the whole (5). Our goal, then, becomes to determine how the entropy of the whole is parcellated out over (potentially multivariate) sharing modes between parts.

P	X_1	X_2
P_{00}	0	0
P_{01}	0	1
P_{10}	1	0
P_{11}	1	1

Table S1. Joint entropy of two discrete random variables that together make up the macro-variable X.

In our toy system given by Table S1, suppose we learn that $X_1 = 0$ OR $X_2 = 0$. Only one global state is excluded: $\mathbf{X} = (1, 1)$ is incompatible with both possibilities, regardless of which is true. Consequently we are only excluding P_{11} from the overall distribution. We can quantify this shared entropy using the *local entropy of shared exclusions* h_{sx}:

$$h^{\mathbf{x}}_{sx}(\{1\}\{2\}) = -\log_2 \mathcal{P}(x_1 \cup x_2) \quad [3]$$

Here, we are adapting the partial entropy notation first introduced by Ince in (15). The function $h^{\mathbf{x}}_{sx}(\{1\}\{2\})$ quantifies the total probability mass of $\mathcal{P}(\mathbf{X})$ excluded by learning either $X_1 = x_1$ or $X_2 = x_2$. Said differently, it is the amount of information that could be learned from either variable alone. Importantly, while it *is* a measure of dependency, it is distinct from the classic mutual information. Unlike the h_{cs} function (1) (and several other redundant mutual information functions, e.g. (7, 9, 15)), our measure h_{sx} operates on the entire joint probability distribution, rather than demanding that redundancy be a function of the pairwise marginals. This sets it apart from other redundancy functions, although it is not unique in this regard (the i_{sx} redundancy function shares this property). For a more detailed discussion of this issue, see Supplementary Material.

So far, we have restricted our examples to the simple case of two variables, x_1 and x_2, however, we are interested in the general case of information common to arbitrarily large, potentially overlapping subsets of a system that has adopted a particular state $\mathbf{x}$. This requires first enumerating the set of subsets, $\mathbf{s}$, which we will call the set of *sources*. It is equivalent to the power set of $\mathbf{x}$, excluding the empty set. For example, if $\mathbf{x} = \{x_1, x_2, x_3\}$, then the source set $\mathbf{s}$ is equal to:

$$\mathbf{s} = \left\{ \begin{matrix} \{x_1\}, \{x_2\}, \{x_3\}, \\ \{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}, \\ \{x_1, x_2, x_3\} \end{matrix} \right\} \quad [4]$$

We are interested in how collections of sources $\mathbf{a} \in \mathbf{s}$ might share entropy (i.e. to what extent they exclude the same possible global configurations of $\mathbf{x}$), which allows us to write our redundant entropy function in full generality. For a collection of sources $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$:

$$h_{sx}(\mathbf{a}_1, \ldots, \mathbf{a}_k) := \log_2 \frac{1}{\mathcal{P}(\mathbf{a}_1 \cup \ldots \cup \mathbf{a}_k)} \quad [5]$$

h_{sx} can be interpreted in terms of logical conjunctions and dysjunctions of variables (12). Consider the example: $h_{sx}(\{x_1\}\{x_2, x_3\})$, which quantifies the amount of probability mass about the state of the whole that would be excluded by observing just the part x_1 **or** the joint state of x_2 **and** x_3. This relationship between probability mass exclusions on one hand, and formal logic on the other, places h_{sx} on a sound conceptual footing. While initially defined locally, it is possible to compute an expected value $\mathcal{H}_{sx}$ for a joint distribution:

$$\mathcal{H}_{sx}(\mathbf{A}_1, \ldots, \mathbf{A}_k) := \mathbb{E}[h_{sx}(\mathbf{a}_1, \ldots, \mathbf{a}_k)] \quad [6]$$

The function h_{sx} is derived from prior work by Makkeh et al. (3), who proposed a redundant information (rather than entropy) function, i_{sx}. This function decomposes the information that a set of sources discloses about a single target, and they noted that if the target was the joint state of all the sources, then the result was a decomposition of the joint entropy of the whole. This is based on the identity that $i(x_1, \ldots, x_k; \mathbf{x}) = h(\mathbf{x})$ (where $\mathbf{x} = \{x_1, \ldots, x_k\}$). Formally:

$$h_{sx}(\mathbf{a}_1, \ldots, \mathbf{a}_k) = i_{sx}(\mathbf{a}_1, \ldots, \mathbf{a}_k; \boldsymbol{\alpha}) \quad [7]$$

This framing makes it intuitively clear that the redundant entropy is the information about the whole that could be learned by observing any of the component parts. While this relationship was noted in (3) and termed h_{sx}, it was not explored in any detail. We have opted to retain their nomenclature (h_{sx}) in our study of the function. For a deeper analysis of the relationship between i_{sx} and h_{sx}, see the Supplementary Material.

B. The Partial Entropy Lattice. Our function h_{sx} has a number of appealing mathematical properties, which collectively satisfy the set of Axioms initially introduced by Williams & Beer for the problem of information decomposition (4) as applied to local information (2, 3):

Symmetry: h_{sx} is invariant under permutation of it's argument: $h_{sx}(\mathbf{a}_1, \ldots, \mathbf{a}_k) = h_{sx}(\sigma(\mathbf{a}_1), \ldots, \sigma(\mathbf{a}_k))$

Monotonicity: h_{sx} decreases as more sources are added: $h_{sx}(\mathbf{a}_1, \ldots, \mathbf{a}_k) \leq h_{sx}(\mathbf{a}_1, \ldots, \mathbf{a}_k, \mathbf{a}_{k+1})$

Self-redundancy: In the special case of a single source, h_{sx} is equivalent to the classic local Shannon entropy: $h_{sx}(\mathbf{a}) = h(\mathbf{a})$.

For proof of these, see (3) Appendix A. Based on these properties, it is possible to specify the domain of h_{sx} (all non-degenerate combinations of sources) in terms of a partially-ordered lattice structure $\mathfrak{A}$ (2, 4). One of the core insights of Williams and Beer was that if one source $\mathbf{a} \subset \mathbf{b}$, then the information about some third source $\mathbf{c}$ redundantly disclosed by $\mathbf{a}$ or $\mathbf{b}$ is equivalent to the information disclosed by $\mathbf{a}$ alone. Consequently, we do not have to compute the redundancy for all possible combination of sources, only all those collections such that no component source is a superset of any other:

$$\mathfrak{A} = \{\boldsymbol{\alpha} \in \mathbb{P}_1(\mathbf{s}) : \forall \mathbf{a}_i, \mathbf{a}_j \in \boldsymbol{\alpha}, \mathbf{a}_i \not\subset \mathbf{a}_j\} \quad [8]$$

Where $\mathbb{P}_1(\mathbf{s})$ indicates the power set of $\mathbf{s}$, excluding the empty set. For an in-depth derivation of the lattice, see (2, 4, 12), for a visualization of the lattice, see Fig. S2. The value of any element $h_\partial(\boldsymbol{\alpha})$ on the lattice can be computed via Mobius inversion:

$$h^{\mathbf{x}}_\partial(\boldsymbol{\alpha}) = h_{sx}(\boldsymbol{\alpha}) - \sum_{\boldsymbol{\beta} \preceq \boldsymbol{\alpha}} h^{\mathbf{x}}_\partial(\boldsymbol{\beta}) \quad [9]$$

The result is the entropy specific to a particular $\boldsymbol{\alpha}$ *and no simpler combination of sources.* Furthermore, the structure of the lattice and the properties of h_{sx} ensure that $h^{\mathbf{x}}_\partial(\boldsymbol{\alpha})$ will always be non-negative. We can re-compute the total joint entropy of $\mathbf{x}$ as:

$$h(\mathbf{x}) = \sum_{i=1}^{|\mathfrak{A}|} h^{\mathbf{x}}_\partial(\boldsymbol{\alpha}_i) \quad [10]$$

Like h_{sx}, it is also possible to compute an expected value of h_∂ (which will also be strictly non-negative):

$$\mathcal{H}^{\mathbf{X}}_\partial(\boldsymbol{\alpha}) = \mathbb{E}[h^{\mathbf{x}}_\partial(\boldsymbol{\alpha})] \quad [11]$$

C. Decomposing Marginal and Joint Entropies. Having defined h_{sx} and the Mobius inversion on the partial entropy lattice, we can now do a complete decomposition of the joint entropy, and its marginal components. For example, consider the bivariate system $\mathbf{X} = \{X_1, X_2\}$. We can decompose the joint entropy:

$$\mathcal{H}(\mathbf{X}) = \mathcal{H}_\partial^{12}(\{1\}\{2\}) + \mathcal{H}_\partial^{12}(\{1\}) + \mathcal{H}_\partial^{12}(\{2\}) + \mathcal{H}_\partial^{12}(\{1,2\}) \quad [12]$$

Furthermore, we can decompose the associated marginal entropies in a manner consistent with the partial information decomposition (4):

$$\mathcal{H}(X_1) = \mathcal{H}_\partial^{12}(\{1\}\{2\}) + \mathcal{H}_\partial^{12}(\{1\}) \quad [13]$$

$$\mathcal{H}(X_2) = \mathcal{H}_\partial^{12}(\{1\}\{2\}) + \mathcal{H}_\partial^{12}(\{2\}) \quad [14]$$

These decompositions can be done for larger ensembles, or more statistical dependencies (see below) and can reveal how higher-order interactions can complicate (and in some cases, compromise) the standard bivariate approaches to functional connectivity.

3. Mathematical Properties of $\mathcal{H}_{sx}$

A. Partial Entropy Decomposition & Partial Information Decomposition. The redundant *entropy* function h_{sx} is closely related to the redundant *information* function i_{sx} proposed by Makkeh et al., (3), and was briefly mentioned in their paper as a possible extension of i_{sx}, although it was not explored in detail, which we will do here. The function h_{sx} is defined:

$$h_{sx}(\mathbf{a}_1, \ldots, \mathbf{a}_k) := \log \frac{1}{\mathcal{P}(\mathbf{a}_1 \cup \ldots \cup \mathbf{a}_k)} \quad [15]$$

This measure is equivalent to the *informative* component of the measure i_{sx} proposed by Makkeh et al., (3) in the context of single-target partial information decomposition. The local redundant information function i_{sx} is defined:

$$i_{sx}(\mathbf{a}^1, \ldots, \mathbf{a}^k; y) := \quad [16]$$

$$\log_2 \frac{\mathcal{P}(y) - \mathcal{P}(y \cap (\bar{\mathbf{a}}^1 \cap \ldots \cap \bar{\mathbf{a}}^k))}{1 - \mathcal{P}(\bar{\mathbf{a}}^1 \cap \ldots \cap \bar{\mathbf{a}}^k)} - \log_2 \mathcal{P}(y) \quad [17]$$

Which can be further decomposed into *informative* and *misinformative* components (3, 5):

$$i_{sx}^+(\mathbf{a}_1, \ldots, \mathbf{a}_k; y) := \log_2 \frac{1}{\mathcal{P}(\mathbf{a}_1 \cup \ldots \cup \mathbf{a}_k)} \quad [18]$$

$$i_{sx}^-(\mathbf{a}_1, \ldots, \mathbf{a}_k; y) := \log_2 \frac{\mathcal{P}(y)}{\mathcal{P}(y \cap (\mathbf{a}_1 \cup \ldots \cup \mathbf{a}_k))} \quad [19]$$

$$i_{sx}(\mathbf{a}_1, \ldots, \mathbf{a}_k; y) = i_{sx}^+(\mathbf{a}_1, \ldots, \mathbf{a}_k; y) - i_{sx}^-(\mathbf{a}_1, \ldots, \mathbf{a}_k; y) \quad [20]$$

Where it is clear that $h_{sx}(\cdot) = i_{sx}^+(\cdot; y)$, with the sole difference that $i_{sx}^+(\cdot; y)$ is implicitly defined with respect to some target variable y (although y has no actual impact on the value). Below, we show that, if the target y is set to the joint state of the whole ($\mathbf{x}$), then the partial entropy decomposition of $h(\mathbf{x})$ with h_{sx} as the shard entropy function becomes equivalent to the partial information decomposition $i(x_1, \ldots, x_N; \mathbf{x})$ with i_{sx} as the redundant entropy function. The notion that the PED is equivalent to doing the PID of the information all the parts disclose about the whole was mentioned parenthetically in (3), although the finding that the informative component is all that is required is novel.

Given the equivalence between $h_{sx}(\cdot)$ and $i_{sx}^+(\cdot; y)$, it suffices to show that $i_{sx}^-(x_1, \ldots, x_N; \mathbf{x}) = 0$ bit in all cases. When $y = \mathbf{x}$, we can re-write the function as:

$$i_{sx}^-(x_1, \ldots, x_N; \mathbf{x}) = \quad [21]$$

$$\log_2 \frac{\mathcal{P}(x_1 \cap \ldots \cap x_N)}{\mathcal{P}((x_1 \cap \ldots \cap x_N) \cap (x_1 \cup \ldots, \cup x_N))}$$

the union of $x^1 \cup \ldots \cup x^k$ is clearly a superset of $x^1 \cap \ldots \cap x^k$, so

$$i_{sx}^- = \log_2 \frac{\mathcal{P}(x_1 \cap \ldots \cap x_N)}{\mathcal{P}(x_1 \cap \ldots \cap x_N)} \quad [22]$$

Which is clearly $\log_2(1) = 0$ bit □

We can understand the partial entropy decomposition using h_{sx} as being equivalent to the decomposition of $i(x_1, \ldots, x_N; \mathbf{x})$. Intuitively, this is consistent with the identity for discrete variables that $\mathcal{I}(\mathbf{X}, \mathbf{X}) = \mathcal{H}(\mathbf{X})$. The vanishing misinformative term can be seen as a special case of the results presented by Ehrlich et al., (16), who proved a general result that the misinformative component of $I_{sx}(\mathbf{X}; Y)$ vanishes if there exists some deterministic function $f : Y \mapsto \mathbf{X}$.

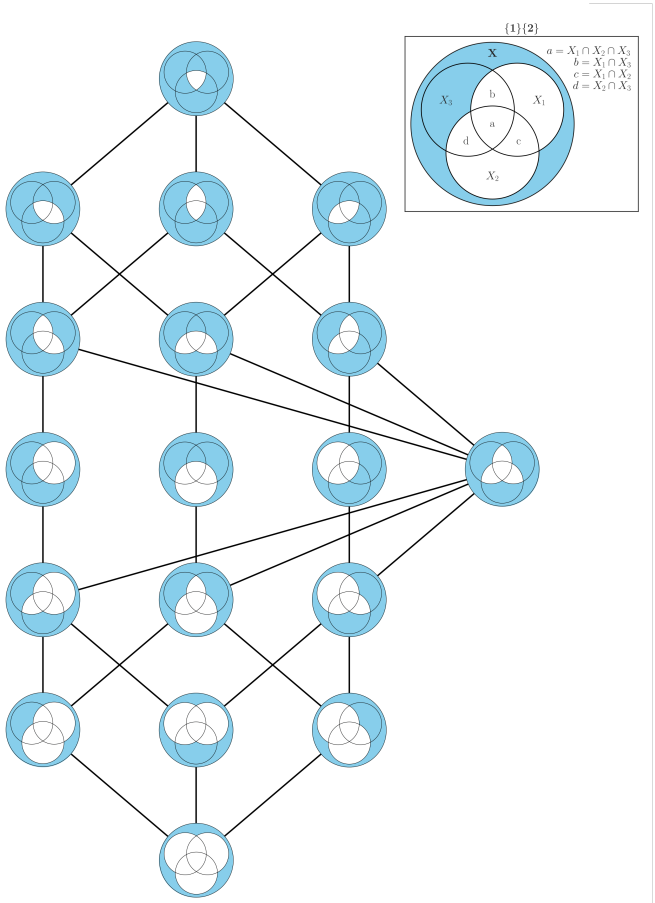

Fig. S2. The partial entropy lattice. The lattice of partial entropy atoms induced by the $\mathcal{H}_{sx}$ function. Each vertex of the lattice corresponds to a single PE atom, and the Venn diagram describes the associated structure of probability mass exclusions. The blue area indicates the probability mass from $\mathcal{P}(\mathbf{x})$ that is excluded by some combination of observations. For example, in the legend, we can see the probability mass excluded by observing $X_1 \vee X_2$. The blue area is all of the probability mass one would exclude after learning the state of *either* component alone. The lowest atom is the entropy redundant to all three elements ($\mathcal{H}_{sx}(\{\{1\}\{2\}\{3\}\})$), and the dependencies get increasingly synergistic higher on the lattice.

XOR					AND				
$\mathcal{P}$	X_1	$\oplus$	X_2	= T	$\mathcal{P}$	X_1	$\wedge$	X_2	= T
1/4	0		0	0	1/4	0		0	0
1/4	0		1	1	1/4	0		1	0
1/4	1		0	1	1/4	1		0	0
1/4	1		1	0	1/4	1		1	1

Table S2. Logical XOR and AND gates.

B. Example: Logical Exclusive-OR (XOR) Gate. To demonstrate how partial entropy decomposition can be used to untangle higher-order interactions, consider the logical exclusive-OR (XOR) gate (for the lookup table, see Table S2). The XOR gate is an example of a *synergistic* logic gate: the ability to predict the state of the target T depends on having access to both X_1 and X_2 jointly: the pariwise marginal mutual informations are equal to 0: $\mathcal{I}(X_1;T) = \mathcal{I}(X_2;T) = 0$ bit, but the joint mutual information is nonzero: $\mathcal{I}(X_1, X_2; T) = 1$ bit.

We can initially see that the triple-redundancy $\mathcal{H}_\partial^{12T}(\{1\}\{2\}\{T\}) = 0$ bit. This is because any configuration of logical disjunctions does not actually rule out any states: for example, $\mathcal{P}(X_1 = 0 \cup X_2 = 0 \cup T = 0) = 1$ as there is no configuration $(1, 1, 1)$ that can be excluded. Other results can be unintuitive. For example, most of the partial entropy is shared between the three bivariate relationships $\mathcal{H}_\partial^{12T}(\{1\}\{2\})$, $\mathcal{H}_\partial^{12T}(\{1\}\{T\})$, and $\mathcal{H}_\partial^{12T}(\{2\}\{T\})$. How is this consistent with the fact that the mutual information between any pair of variables is zero? The bivariate redundancy can be non-zero in this case because, on average, knowing the local state of $x_1 \vee x_2$ reduces our uncertainty about the joint state of $\{x_1, x_2, t\}$. For example, suppose we learn that $x_1 = 1 \vee x_2 = 1$. This *excludes* the joint configuration $\{x_1 = 0, x_2 = 0, t = 0\}$. This exclusion of the associated probability mass is recognized by $h_{sx}(\cdot)$ as informative, in that it reduces our uncertainty about the joint-state of the whole, despite the fact that, on average, X_1 and X_2 disclose no information about T. There is no redundant information common to X_1, X_2 and T, however, and there a number of higher-order dependencies, such as $\mathcal{H}_\partial^{12T}(\{1\}\{2, T\})$ and $\mathcal{H}_\partial^{12T}(\{1, 2\}\{1, T\}\{2, T\})$.

Atom	$\mathcal{H}_\partial^{12T}\|\|$	XOR	AND	MaxEnt
$\{1\}\{2\}\{T\}$		0.0	0.208	0.193
$\{1\}\{2\}$		0.415	0.208	0.222
$\{1\}\{T\}$		0.415	0.25	0.222
$\{2\}\{T\}$		0.415	0.25	0.222
$\{1\}\{2, T\}$		0.17	0.04	0.041
$\{2\}\{1, T\}$		0.17	0.04	0.041
$\{T\}\{1, 2\}$		0.17	0.104	0.041
$\{1\}$		0.0	0.292	0.322
$\{2\}$		0.0	0.292	0.322
$\{T\}$		0.0	0.0	0.322
$\{1, 2\}\{1, T\}\{2, T\}$		0.245	0.0	0.018
$\{1, 2\}\{1, T\}$		0.0	0.104	0.093
$\{1, 2\}\{2, T\}$		0.0	0.104	0.093
$\{1, T\}\{2, T\}$		0.0	0.0	0.093
$\{1, 2\}$		0.0	0.104	0.17
$\{1, T\}$		0.0	0.0	0.17
$\{2, T\}$		0.0	0.0	0.17
$\{1, 2, T\}$		0.0	0.0	0.245

Table S3. The Partial Entropy Decomposition for the XOR, AND, and Maximum Entropy Gates.

C. Independent Variables. One unusual property of h_{sx}, as demonstrated by the logical-XOR results is that independent variables can still share entropy. This is a recognized feature of multiple measures of redundant information/entropy and is generally considered to be an issue to be excised (14) (some have gone so far as the suggest an axiom that such a property must be disallowed from the outset (15)). While we understand that shared entropy for random variables may seem initially counter intuitive, it can be readily understood when considering the problem of inference. Let us return to our two element example (Table 1), and this time specify that $X_1 \perp X_2$. We know then that $\mathcal{I}(X_1; X_2) = 0$ bit, however, $h_{sx}(\{X_1\}\{X_2\}) \approx 0.415$ bit. Why? The answer is that, while the two variables are independent, in all cases learning either $X_1 = x_1 \vee X_2 = x_2$ is sufficient to exclude a single possible state: the case where $X_1 = \neg x_1 \wedge X_2 = \neg x_2$. If we were to formalize this in terms of a gambling problem, we would find that, despite the independence of both variables, a player is, in fact, more likely to win with a correct guess after learning $X_1 \vee X_2$. See Figure S3.

Furthermore, we can see that, while h_{sx} will be greater than zero for small, maximum entropy systems, as the system gets larger, the redundancy will logarithmically trend towards zero. The proof for binary systems is straightforward. For a discrete, maximum entropy system with k elements, learning the state of $X_1 = x_1 \vee \ldots \vee X_k = x_k$ will always exclude a single state: the state where $X_1 \neq x_1 \wedge \ldots \wedge X_k \neq x_k$. This single state $\mathbf{x}^*$ will have $\mathcal{P}(\mathbf{x}^*) = 1/k$ (as all states have the same probability by the maximum entropy constraint). The union of all surviving configurations will be $1 - \mathcal{P}(\mathbf{x}^*)$. Since $\lim_{k\to\infty} 1/k = 0$, then the

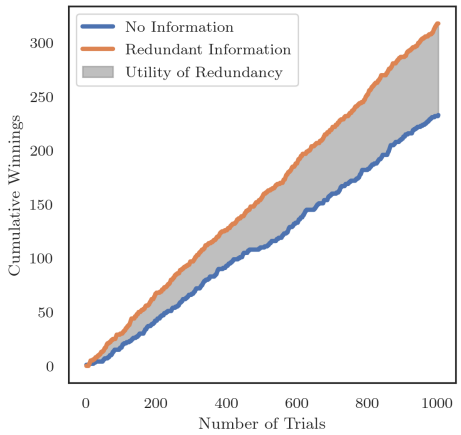

Fig. S3. Utility of redundant information. Suppose an agent plays a gambling game, where two independent, binary variables are set at random (so all outcomes $\mathcal{P}(x_1, x_2) = 1/4$ for all configurations). If the agent guesses the correct configuration of both variables, they win \$1 and if they guess wrong, they win nothing. Clearly, the expected value of each trial is \$0.25 (blue curve). However, if another agent learns that $X_1 = x1 \vee X_2 = x2$, then they can do better at the game, with an expected value of each trial of \$0.33. The difference between the two cumulative distributions of 1000 trials is the extra value that can be extracted from the redundant information. This shows that, while counter-intuitive, the fact that $\mathcal{H}_\partial^{12}(\{1\}\{2\}) > 0$ even if $X_1 \perp X_2$ is interpretable in practical contexts.

union probability will $\to 1$ and consequently $h_{sx} \to 0$ bit. This suggests that, for very large, idealized systems (such as an ideal gas), the redundancy *does* go to 0 bit for maximum entropy systems. How other values (such as the redundant and synergistic structure) behave remains an area of further study, although we conjecture that, as $k \to \infty$, redundancies and synergies will vanish faster than unique terms.

D. $\mathcal{H}_{\mathbf{sx}}$ & Pairwise Dependency Constraints. The h_{sx} measure is built on the full joint distribution, rather than the maximum entropy distribution that preserves pairwise marginals. While this as a departure from historical approaches to PID, we do not consider it necessarilly problematic for two reasons:

In Bertschinger's original proposal (7, 8) that the redundancy depends solely on pairwise marginals, the focus was on the particular case of two inputs X_1, X_2 that were jointly disclosing information about a single target Y. In this context, the focus on pairwise marginals makes sense, as the double synergy term $\{X_1, X_2\}$ is the only relevant higher-order term. However, as the number of elements grows, we feel the focus on pairwise marginals becomes less natural.

For instance, consider the atom $\{1, 2, 3\}\{2, 3, 4\}$, which appears in the partial entropy lattice for a set of four elements $\{X_1, X_2, X_3, X_4\}$. Since this atom only contains higher-order information in the joint states of three variables, it does not feel natural that it must be computed from the marginals X_1, X_2, etc. At this point on the lattice, it is assumed that the observer has access to the joint states of at least three variables: (X_1, X_2, X_3) and the joint state of (X_2, X_3, X_4). Based on Bertschinger et al. (7, 8), it could be argued that the atom $\{1, 2, 3\}\{2, 3, 4\}$ should be calculated with respect to the distribution that preserves third-order marginals, however that opens a difficult can of worms: should every atom in the lattice be computed with respect to the maximum entropy distribution that preserves the marginals of whatever the lowest-order source in the atoms is? While this is an intriguing possibility, attempting to implement this is beyond the existing capabilities of our code and the `SxPID` package.

There is another angle as well. If we consider the interpretation that $h_{sx}(x_1, x_2, x_3) = i_{sx}(x_1, x_2, x_3; \mathbf{x})$ where $\mathbf{x} = \{x_1, x_2, x_3\}$, it seems as though knowledge of the whole is implicitly built in, even when considering redundant atoms such as $\{1\}\{2\}\{3\}$. If we follow the logic that redundancy should only depend on the pairwise dependency between each input and the target, since $\mathbf{X}$ is the target, the joint state of the whole is already accounted for.

Ince's PED approach (1) required computing h_{ccs} with respect to the maximum entropy distribution that preserves pairwise marginals. In this case though, since the distribution of the whole is the target distribution, this approach seems unnatural, since involves changing the statistics of the whole: effectively decomposing the entropy of an entirely different system than the one under study. Ultimately, h_{sx} as it is currently formulated, cannot be easily be reconfigured to accommodate the pairwise dependency restriction.

4. Basic Information Theory Review

Here we will provide a basic overview of information theory for unfamiliar readers. For a more comprehensive treatment of the subject, see the textbooks by Cover & Thomas (10) and/or MacKay (11).

The basic object of study in information theory is the *entropy*, which quantifies the total uncertainty that we, as observers, have about the state of some variable X. For the purposes of this paper, we will assume that X is discrete, with a finite number of possible states that can be pulled from the support set $\mathcal{X}$. For every particular state $x \in \mathcal{X}$, there is an associated probability $\mathcal{P}(x)$. The entropy of X is given by:

$$H(X) = -\sum_{x \in \mathcal{X}} \mathcal{P}(x) \log \mathcal{P}(x) \quad [23]$$

For multiple variables, we can define the joint entropy as:

$$H(X_1, X_2) = -\sum_{\substack{x_1 \in \mathcal{X}_1 \\ x_2 \in \mathcal{X}_2}} \mathcal{P}(x_1, x_2) \log \mathcal{P}(x_1, x_2) \quad [24]$$

We can also define the conditional entropy as the uncertainty about X_1 *left over* after accounting for the knowledge that $X_2 = x_2$:

$$H(X_1|X_2) = -\sum_{\substack{x_1 \in \mathcal{X}_1 \\ x_2 \in \mathcal{X}_2}} \mathcal{P}(x_1, x_2) \log \mathcal{P}(x_1|x_2) \quad [25]$$

From these basic components, we can define the *mutual information* as the difference between our initial uncertainty about the state of X_1 the the remaining uncertainty about X_1 that is not resolved by learning the state of X_2:

$$I(X_1; X_2) = H(X_1) - H(X_1|X_2) \quad [26]$$

The mutual information is symmetric in it's arguments: $I(X_1; X_2) = I(X_2; X_1)$. If we have multiple Xs disclosing information about a single target T, the joint mutual information has the same form:

$$I(X_1, X_2; T) = H(T) - H(T|X_1, X_2) \quad [27]$$

The mutual information can also be written in terms of probabilities:

$$I(X_1; X_2) = \sum_{\substack{x_1 \in \mathcal{X}_1 \\ x_2 \in \mathcal{X}_2}} \mathcal{P}(x_1, x_2) \log \frac{\mathcal{P}(x_1|x_2)}{\mathcal{P}(x_1)} \quad [28]$$

A. Local Information Theory. Both the entropy and the mutual information can be understood as expected values over some (potentially multivariate) distribution):

$$H(X) = \mathbb{E}[-\log \mathcal{P}(x)] \quad [29]$$

The term $-\log \mathcal{P}(x)$ is known as the *local entropy* or the *Shannon information content* and it quantifies how surprised we, as observers are to see that $X = x$. It is typically denoted as $h(x)$.

$$I(X_1; X_2) = \mathbb{E}\left[\log_2 \frac{\mathcal{P}(x_1|x_2)}{\mathcal{P}(x_1)}\right] \quad [30]$$

The term $\frac{\mathcal{P}(x_1|x_2)}{\mathcal{P}(x_1)}$ is known as the *local mutual information* and it quantifies the divergence between the prior probability $X_1 = x_1$ and the posterior probability $X_1 = x_1$ *after accounting for the fact that* $X_2 = x_2$. It is typically denoted as $i(x_1; x_2)$. Unlike the expected mutual information, which is strictly non-negative, the local mutual information can be either greater than, or less than, zero. If $\mathcal{P}(x_1|x_2) < \mathcal{P}(x_1)$, then $i(x_1; x_2) > 0$, and if $\mathcal{P}(x_1|x_2) < \mathcal{P}(x_1)$, then $i(x_1; x_2) < 0$. In the latter case, we say that x_1 *misinforms* on the state of x_2.

5. Derivations

For didactic purposes, we have included a number of derivations of the relationships between mutual information and partial entropy atoms. The basic logic is reasonably straightforward: any information-theoretic construct that can be written in terms of joint and marginal entropies can be converted into partial entropy atoms by first decomposing each of the constituent entropies, and then summing them together in accordance with the original definition. Importantly, all atoms must be pulled from the lattice describing the structure of the whole system.

A. Derivation of Eq. 5. Eq. 5 shows:

$$\mathcal{I}(X_1; X_2) = \mathcal{H}_\partial^{12}(\{1\}\{2\}) - \mathcal{H}_\partial^{12}(\{1, 2\})$$

This can be derived from:

$$\mathcal{I}(X_1; X_2) = \mathcal{H}(X_1) + \mathcal{H}(X_2) - \mathcal{H}(X_1, X_2)$$

From Eqs. 2, 3 and 4 we have:

$$\begin{aligned}\mathcal{H}(X_1; X_2) &= \mathcal{H}_\partial^{12}(\{1\}\{2\}) + \mathcal{H}_\partial^{12}(\{1\}) \\ &\quad + \mathcal{H}_\partial^{12}(\{2\}) + \mathcal{H}_\partial^{12}(\{1,2\}) \\ \mathcal{H}(X_1) &= \mathcal{H}_\partial^{12}(\{1\}\{2\}) + \mathcal{H}_\partial^{12}(\{1\}) \\ \mathcal{H}(X_2) &= \mathcal{H}_\partial^{12}(\{1\}\{2\}) + \mathcal{H}_\partial^{12}(\{2\})\end{aligned}$$

Basic substitution shows that:

$$\begin{aligned}\mathcal{I}(X_1; X_2) = \mathcal{H}_\partial^{12}(\{1\}\{2\}) + \mathcal{H}_\partial^{12}(\{1\}) + \mathcal{H}_\partial^{12}(\{1\}\{2\}) + \mathcal{H}_\partial^{12}(\{2\}) \\ -\mathcal{H}_\partial^{12}(\{1\}\{2\}) - \mathcal{H}_\partial^{12}(\{1\}) - \mathcal{H}_\partial^{12}(\{2\}) - \mathcal{H}_\partial^{12}(\{1,2\})\end{aligned}$$

Simplifying returns Eq. 5.

B. Derivation of Eq. 6. Eq. 16 shows that in a triad X_1, X_2, X_3, the bivariate mutual information decomposes as:

$$\begin{aligned}\mathcal{I}(X_1; X_2) = \mathcal{H}_\partial^{123}(\{1\}\{2\}\{3\}) + \mathcal{H}_\partial^{123}(\{1\}\{2\}) \\ - \mathcal{H}_\partial^{123}(\{3\}\{1,2\}) - \mathcal{H}_\partial^{123}(\{1,2\}\{1,3\}\{2,3\}) \\ - \mathcal{H}_\partial^{123}(\{1,2\}\{1,3\}) - \mathcal{H}_\partial^{123}(\{1,2\}\{2,3\}) \\ - \mathcal{H}_\partial^{123}(\{1,2\})\end{aligned}$$

The logic is essentially the same as was given above, however, atoms are drawn from the three element lattice:

$$\begin{aligned}\mathcal{H}(X_1) &= \mathcal{H}_\partial^{123}(\{1\}\{2\}\{3\}) + \mathcal{H}_\partial^{123}(\{1\}\{2\}) + \mathcal{H}_\partial^{123}(\{1\}\{3\}) \\ &\quad + \mathcal{H}_\partial^{123}(\{1\}\{2,3\}) + \mathcal{H}_\partial^{123}(\{1\}) \\ \mathcal{H}(X_2) &= \mathcal{H}_\partial^{123}(\{1\}\{2\}\{3\}) + \mathcal{H}_\partial^{123}(\{1\}\{2\}) + \mathcal{H}_\partial^{123}(\{2\}\{3\}) \\ &\quad + \mathcal{H}_\partial^{123}(\{2\}\{1,3\}) + \mathcal{H}_\partial^{123}(\{2\}) \\ \mathcal{H}(X_1, X_2)) &= \mathcal{H}_\partial^{123}(\{1\}\{2\}\{3\}) + \mathcal{H}_\partial^{123}(\{1\}\{2\}) + \mathcal{H}_\partial^{123}(\{1\}\{3\}) + \mathcal{H}_\partial^{123}(\{2\}\{3\}) \\ &\quad + \mathcal{H}_\partial^{123}(\{1\}\{2,3\}) + \mathcal{H}_\partial^{123}(\{2\}\{1,3\}) + \mathcal{H}_\partial^{123}(\{3\}\{1,2\}) \\ &\quad + \mathcal{H}_\partial^{123}(\{1\}) + \mathcal{H}_\partial^{123}(\{2\}) + \mathcal{H}_\partial^{123}(\{1,2\}\{1,3\}\{2,3\}) \\ &\quad + \mathcal{H}_\partial^{123}(\{1,2\}\{1,3\}) + \mathcal{H}_\partial^{123}(\{1,2\}\{2,3\}) \\ &\quad + \mathcal{H}_\partial^{123}(\{1,2\})\end{aligned}$$

Once again, substitution and simplification shows:

$$\begin{aligned}\mathcal{I}(X_1; X_2) = \mathcal{H}_\partial^{123}(\{1\}\{2\}\{3\}) + \mathcal{H}_\partial^{123}(\{1\}\{2\}) \\ - \mathcal{H}_\partial^{123}(\{3\}\{1,2\}) - \mathcal{H}_\partial^{123}(\{1,2\}\{1,3\}\{2,3\}) \\ - \mathcal{H}_\partial^{123}(\{1,2\}\{1,3\}) - \mathcal{H}_\partial^{123}(\{1,2\}\{2,3\}) \\ - \mathcal{H}_\partial^{123}(\{1,2\})\end{aligned}$$

Similar logic underlies the derivation of Eq. 7, using the identity that $\mathcal{I}(X_1; X_2|X_3) = \mathcal{H}(X_1; X_3) + \mathcal{H}(X_2; X_3) - \mathcal{H}(X_1, X_2, X_3) - \mathcal{H}(X_3)$.

C. Derivation of Eq. 10. Equation 10 describes how the total correlation can be decomposed into partial entropy atoms:

$$\begin{aligned}\mathcal{T}(X_1, X_2, X_3) = (2 \times \mathcal{H}_\partial^{123}(\{1\}\{2\}\{3\})) \\ + \mathcal{H}_\partial^{123}(\{1\}\{2\}) + \mathcal{H}_\partial^{123}(\{1\}\{3\} + \{2\}\{3\}) \\ - \mathcal{H}_\partial^{123}(\{1,2\}\{1,3\}\{2,3\}) \\ - \mathcal{H}_\partial^{123}(\{1,2\}\{1,3\}) - \mathcal{H}_\partial^{123}(\{1,2\}\{2,3\}) - \mathcal{H}_\partial^{123}(\{1,3\}\{2,3\}) \\ - \mathcal{H}_\partial^{123}(\{1,2\}) - \mathcal{H}_\partial^{123}(\{1,3\}) - \mathcal{H}_\partial^{123}(\{2,3\}) \\ - \mathcal{H}_\partial^{123}(\{1,2,3\})\end{aligned}$$

The logic of the decomposition is the same as above: we begin by writing the total correlation out in terms of joint and marginal entropies:

$$\mathcal{T}(X_1, X_2, X_3) := \sum_{i=1}^{3} \mathcal{H}(X_i) - \mathcal{H}(X_1, X_2, X_3)$$

which can be decomposed:

$$\begin{aligned}
\mathcal{H}(X_1) &= \mathcal{H}_\partial^{123}(\{1\}\{2\}\{3\}) + \mathcal{H}_\partial^{123}(\{1\}\{2\}) + \mathcal{H}_\partial^{123}(\{1\}\{3\}) \\
&\quad + \mathcal{H}_\partial^{123}(\{1\}\{2,3\}) + \mathcal{H}_\partial^{123}(\{1\}) \\
\mathcal{H}(X_2) &= \mathcal{H}_\partial^{123}(\{1\}\{2\}\{3\}) + \mathcal{H}_\partial^{123}(\{1\}\{2\}) + \mathcal{H}_\partial^{123}(\{2\}\{3\}) \\
&\quad + \mathcal{H}_\partial^{123}(\{2\}\{1,3\}) + \mathcal{H}_\partial^{123}(\{2\}) \\
\mathcal{H}(X_3) &= \mathcal{H}_\partial^{123}(\{1\}\{2\}\{3\}) + \mathcal{H}_\partial^{123}(\{1\}\{3\}) + \mathcal{H}_\partial^{123}(\{2\}\{3\}) \\
&\quad + \mathcal{H}_\partial^{123}(\{3\}\{1,2\}) + \mathcal{H}_\partial^{123}(\{3\}) \\
\mathcal{H}(X_1, X_2)) &= \mathcal{H}_\partial^{123}(\{1\}\{2\}\{3\}) + \mathcal{H}_\partial^{123}(\{1\}\{2\}) + \mathcal{H}_\partial^{123}(\{1\}\{3\}) + \mathcal{H}_\partial^{123}(\{2\}\{3\}) \\
&\quad + \mathcal{H}_\partial^{123}(\{1\}\{2,3\}) + \mathcal{H}_\partial^{123}(\{2\}\{1,3\}) + \mathcal{H}_\partial^{123}(\{3\}\{1,2\}) \\
&\quad + \mathcal{H}_\partial^{123}(\{1\}) + \mathcal{H}_\partial^{123}(\{2\}) + \mathcal{H}_\partial^{123}(\{3\}) + \mathcal{H}_\partial^{123}(\{1,2\}\{1,3\}\{2,3\}) \\
&\quad + \mathcal{H}_\partial^{123}(\{1,2\}\{1,3\}) + \mathcal{H}_\partial^{123}(\{1,2\}\{2,3\}) + \mathcal{H}_\partial^{123}(\{1,2\}\{2,3\}) \\
&\quad + \mathcal{H}_\partial^{123}(\{1,2\}) + \mathcal{H}_\partial^{123}(\{1,3\}) + \mathcal{H}_\partial^{123}(\{2,3\}) \\
&\quad + \mathcal{H}_\partial^{123}(\{1,2,3\})
\end{aligned}$$

The same logic can be applied to Eq. 12 to decompose the dual total correlation (with the insight that $\mathcal{H}(X_1|X_2, X_3) = \mathcal{H}(X_1, X_2, X_3) - \mathcal{H}(X_2, X_3)$), and from there, the O-information can be computed via subtraction of the sets of atoms.

6. Materials & Methods

A. Human Connectome Project fMRI Data. The data used in this study was taken from a set of 100 unrelated subjects included in the Human Connectome Project (HCP) (17). Refs (17, 18) provide a detailed description of the acquisition and preprocessing of this data, which have been used in many previous studies (19, 20). Briefly, all subjects gave informed consent to protocols approved by the Washington University Institutional Review Board. Data was collected with a Siemens 3T Connectom Skyra using a head coil with 32 channels. Functional data analysed here was acquired during resting state with a gradient-echo echo-planar imaging (EPI) sequence. Collection occurred over four scans on two separate days (scan duration: 14:33 min; eyes open). The main acquisition parameters included TR = 720 ms, TE = 33.1 ms, flip angle of 52°, 2 mm isotropic voxel resolution, and a multiband factor of 8. Resting state data was mapped to a 200-node parcellation scheme (21) covering the entire cerebral cortex.

Considerations for subject inclusion were established before the study and are as follows. The mean and mean absolute deviation of the relative root mean square (RMS) motion throughout any of the four resting scans were calculated. Subjects that exceeded 1.5 times the interquartile range in the adverse direction for two or more measures they were excluded. This resulted in the exclusion of four subjects, and an additional subject due to a software error during diffusion MRI processing. The included subjects had demographic characteristics of: 56% female, mean age = 29.29 ± 3.66, age range = 22-36 years.

A.1. Preprocessing. The minimal preprocessing of HCP rs-fMRI data can be found described in detail in Ref. (18). Five main steps were followed: 1) susceptibility, distortion, and motion correction; 2) registration to subject-specific T1-weighted data; 3) bias and intensity normalization; 4) projection onto the 32k_fs_LR mesh; and 5) alignment to common space with a multimodal surface registration (81). This pipeline produced an ICA+FIX time series in the CIFTI grayordinate coordinate system. We included two additional preprocessing steps: 6) global signal regression and 7) detrending and band pass filtering (0.008 to 0.08 Hz) (22). We discarded the first and last 50 frames of each time series after confound regression and filtering to produce final scans with length 13.2 min (1,100 frames). All four scans from 95 subjects were then z-scored and concatenated to give a final time-series of 200 brain regions and 418,000 time points.

A.2. Discretizing BOLD Signals. Unfortunately, the $\mathcal{H}_{sx}$ measure is only well-defined for discrete random variables. Consequently, we discretized our data by binarizing the z-scored time series: setting any value greater than zero to one and any value less than zero to zero. Prior work has established that transforming BOLD signals into binary point processes preserves the majority of the total correlation structure (19, 23), so we are confident that our analysis is robust, especially considering the large number of samples.

We chose to binarize around the z-score (as opposed to alternative point-processing techniques such as local maxima), as the z-score ensures that each individual channel is generally maximally entropic (i.e. $\mathcal{P}(X_i = 1) \approx \mathcal{P}(X_i = 0) \approx 1/2$). This ensures that every individual channel has approximately the same entropy, and so deviations from maximum entropy at the

level of the entire triad or tetrad can *only* emerge from correlations between two or more channels, rather than being influenced by biases at the channel-level. The choice to binarize about the mean also links this work to previous work on decomposing functional connectivity into discrete partitions (19).

B. Statistical Analyses.

B.1. Triads & tetrads. In standard FC analysis, it is typical to compute the pairwise correlation between all pairs of brain regions, resulting $\binom{N}{2}$ unique pairs. For this analysis, we computed all triads of brain regions, resulting in $\binom{200}{3} = 1,313,400$ unique triples. For each triad, we computed the joint entropy, and performed the full partial entropy decomposition to compute each of the eighteen partial entropy atoms. Finally, each of the atoms was normalized by the total joint entropy, to give a measure of how much each atom contributes to the whole entropy. This allows us to directly compare triads that have different joint entropies.

It was not feasible to brute-force all possible tetrads, which is a set of approximately sixty-four million. Instead, we randomly sub-sampled sets of four randomly, collecting 1954000 tetrads ($\approx 3\%$ of the total space) and analyzing them.

B.2. Bivariate functional connectivity networks. To directly compare the PED framework to the standard, correlation-based FC network framework, we constructed single, representative FC network by computing the pairwise mutual information between every pair of regions in the fMRI scan (as was done in (20)).

$$\mathcal{I}(X;Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X,Y) \quad [31]$$

B.3. Subgraph Analysis. Since we are interested in how the bivariate FC framework reflects (or fails to reflect) higher-order redundancies and synergies, we also compute a battery of structure metrics on matching subgraphs taken from the FC network. Formally presented by Onnela et al., (13), we consider arithmetic mean of the subgraph connectivity:

$$\mathcal{G}_{\mathfrak{A}}(\mathbf{X}) = \frac{\sum_{i \neq j} \mathcal{I}(X_i; X_j)}{|\mathbf{X}|^2 - |\mathbf{X}|} \quad [32]$$

For a given triad of tetrad $\mathbf{X}$, we compared the mean FC density to the various redundant and synergistic information-sharing structures of $\mathbf{X}$.

B.4. Community Detection on Bivariate Matrices. Multi-resolution consensus clustering (26) was used to detect network communities in the functional connectivity matrix across multiple scales. The algorithm proceeds in three main stages. In the first stage, modularity maximization using the Louvain method is performed for 1,000 different values of the resolution parameter, γ. This produced a range of γ values that resulted with partitions having between 2 and N communities. The second stage consisted of a more fine-grained sweep (10,000 steps) over the γ values defined in the first stage of the process. We aggregate the partitions produced by this sweep into a node-by-node co-classification matrix storing how frequently nodes are partitioned into the same community. A null model with expected values of co-classification based on the size and number of communities was subtracted from the co-classification matrix (26). Finally, in the third stage, the null-adjusted co-classification matrix was clustered again using consensus clustering with 100 repetitions and a consensus threshold τ of 0 (24). The resulting partition was used for analyses.

We assessed the similarity between single-subject partitions and consensus partitions using Normalized Mutual Information (NMI). Each partition can be formalized as a vector of integers of dimension N whose entries denote the nodes' allegiance to communities. NMI estimates the similarity between two partitions by counting co-occurrences in the two vectors.

We computed NMI between each one of the 95 single-subject partitions and the consensus partition, in both cases of redundancy and synergy hypergraphs. We assessed the significance of NMI values by comparing them with a null case obtained by randomly shuffling 1000 times communities labels in the single-subject partitions. The p-values of the statistical test, calculated as the fraction of null-case NMI greater than the actual NMI, have been corrected with a Benjamini-Hochberg test.

B.5. Null Model. To ensure that the statistical dependencies we were observing reflect non-trivial interactions, we significance-tested triads and tetrads against a null distribution composed of one million, maximum entropy null models. We constructed sets of totally independent, maximum entropy binary time series and computed the PED on each set of three or four null channels. From this, we can construct distributions of the expected null structure and expected synergistic structure against which to compare triads and tetrads.

B.6. Hypergraph Community Detection. Each of the triads can be thought of as a hyper-edge on a 3-uniform hypergraph of 200 nodes. For the synergistic structure, we selected only those hyperedges who had a *greater* synergistic structure than any of the one million maximum-entropy nulls that formed our null distribution. This resulted in a hypergraph with 200 hundred nodes and 3,746 regular hyper-edges. We used the same criteria to build a redundant structure hypergraph using the top 3,746 most redundant hyperedges.

Both hypergraphs were clustered using the `HyperNetX` package (available on Github: https://github.com/pnnl/HyperNetX) implementation of the hyper-modularity optimization by Kumar and Vaidyanathan et al., (27).

Briefly, the algorithm by Kumar and Vaidyanathan et al., takes a modularity maximization approach to partitioning the vertices of a hypergraph into non-overlapping communities. In dyadic networks, the modularity function compares the

distribution of within- and between-community edges to the expected distribution based on a degree-preserving, configuration null model (25). In the case of hypergraphs, a hyper-configuration model can be used instead. A generalized modularity metric can then be used as an objective function in a Louvain-based, modularity maximization search.

B.7. Temporal Structure. To explore the temporal structure of the state-transition series, we used the active information storage (28, 29) (a measure of how predictable is the future given the past) and the determinism (30, 31), (a measure of how constrained the future is given the past). For a one dimensional, discrete random variable X that evolves through time, we can compute the information that the past X_{t-1} discloses about the future X_t with the mutual information:

$$\mathcal{AIS}(X) = \mathcal{I}(X_{t-1}; X_t) \quad [33]$$

This measure quantifies the degree to which knowing the past reduces our uncertainty about the future. This term can be further decomposed into two components: the determinism and the degeneracy (30):

$$\mathcal{I}(X_{t-1}; X_t) = Det(X) - Deg(X) \quad [34]$$

Where determinism is:

$$Det(X) = \log_2(N) - \mathcal{H}(X_t|X_{t-1}) \quad [35]$$

And degeneracy is:

$$Deg(X) = \log_2(N) - \mathcal{H}(X_t) \quad [36]$$

The determinism quantifies how reliably a given past state x_{t-1} leads to a single future state x_t. If $\mathcal{P}(x_t|x_{t-1}) \approx 1$, then we say that x_{t-1} *deterministically* leads to x_t.

We significance tested both the active information storage and the determinism by comparing the empirical values to an ensemble of ten thousand randomly permuted nulls generated by shuffling the time series. Since the degeneracy is unchanged by permutation of the temporal structure (since the marginal entropy $H(X_t)$ is the same), any changes in active information storage produced by shuffling must be driven by changes in the determinism.

C. Software. All partial information/entropy decompositions were done using the `SxPID` package released with (3) and can be accessed on Github: https://github.com/Abzinger/SxPID.

D. Data & Software Sharing. Data and software are provided as supplementary material. Supplementary software includes:

- **Code S1**: Cython script for computing PEDs for triads and tetrads. Change the .txt extension to .pyx.
- **Code S2**: Python code for analysing the rank-differences between triads, the framewise similarities, and the active information storage analysis. Change the .txt extension to .py.
- **Code S3**: Python code for analysing the relationship between PEDs and bivariate correlations. Constructs Figure 1. Change the .txt extension to .py.
- **Code S4**: Python code for constructing Figure 2 and 3. Change the .txt extension to .py.

Supplementary datasets include:

- **Datasets S1-10**: 10 .csv files containing results for the triads.
- **Dataset S11-20**: 10 .csv files containing results for the tetrads.
- **Dataset S21-30**: 10 .csv files containing the null triad analyses.

References

1. Robin A. A. Ince. The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. *arXiv:1702.01591 [cs, math, q-bio, stat]*, February 2017. arXiv: 1702.01591.
2. Conor Finn and Joseph T. Lizier. Pointwise Partial Information Decomposition Using the Specificity and Ambiguity Lattices. *Entropy*, 20(4):297, April 2018. Number: 4
3. Abdullah Makkeh, Aaron J. Gutknecht, and Michael Wibral. Introducing a differentiable measure of pointwise shared information. *Physical Review E*, 103(3):032149, March 2021.
4. Paul L. Williams and Randall D. Beer. Nonnegative Decomposition of Multivariate Information. *arXiv:1004.2515 [math-ph, physics:physics, q-bio]*, April 2010. arXiv: 1004.2515.
5. Conor Finn and Joseph T. Lizier. Probability Mass Exclusions and the Directed Components of Mutual Information. *Entropy*, 20(11):826, November 2018. Number: 11
6. Conor Finn and Joseph T. Lizier. Generalised Measures of Multivariate Information Content. *Entropy*, 22(2):216, February 2020. Number: 2

7. Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying Unique Information. *Entropy*, 16(4):2161–2183, April 2014. Number: 4
8. Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, and Jürgen Jost. Shared Information – New Insights and Problems in Decomposing Information in Complex Systems arXiv:1210.5902 [cs, math].
9. Adam B. Barrett. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Physical Review E*, 91(5):052802, May 2015.
10. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, November 2012.
11. David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, September 2003.
12. A. J. Gutknecht, M. Wibral, and A. Makkeh. Bits and pieces: understanding information decomposition from part-whole relationships and formal logic. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2251):20210110, July 2021.
13. Jukka-Pekka Onnela, Jari Saramäki, János Kertész, and Kimmo Kaski. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6):065103, June 2005.
14. Artemy Kolchinsky. A Novel Approach to the Partial Information Decomposition. *Entropy*, 24(3):403, March 2022. Number: 3
15. Robin A. A. Ince. Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. *Entropy*, 19(7):318, July 2017. Number: 7
16. David A. Ehrlich, Andreas C. Schneider, Michael Wibral, Viola Priesemann, and Abdullah Makkeh. Partial Information Decomposition Reveals the Structure of Neural Representations, September 2022. arXiv:2209.10438 [cs, math, stat].
17. David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E. J. Behrens, Essa Yacoub, and Kamil Ugurbil. The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80:62–79, October 2013.
18. Matthew F. Glasser, Stamatios N. Sotiropoulos, J. Anthony Wilson, Timothy S. Coalson, Bruce Fischl, Jesper L. Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R. Polimeni, David C. Van Essen, Mark Jenkinson, and WU-Minn HCP Consortium. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80:105–124, October 2013.
19. Olaf Sporns, Joshua Faskowitz, Andreia Sofia Teixeira, Sarah A. Cutts, and Richard F. Betzel. Dynamic expression of brain functional systems disclosed by fine-scale analysis of edge time series. *Network Neuroscience*, 5(2):405–433, May 2021.
20. Maria Pope, Makoto Fukushima, Richard F. Betzel, and Olaf Sporns. Modular origins of high-amplitude cofluctuations in fine-scale functional connectivity dynamics. *Proceedings of the National Academy of Sciences*, 118(46), November 2021.
21. Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Avram J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex (New York, N.Y.: 1991)*, 28(9):3095–3114, September 2018.
22. Linden Parkes, Ben Fulcher, Murat Yücel, and Alex Fornito. An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. *NeuroImage*, 171:415–436, May 2018.
23. Enzo Tagliazucchi, Pablo Balenzuela, Daniel Fraiman, and Dante R. Chialvo. Criticality in Large-Scale Brain fMRI Dynamics Unveiled by a Novel Point Process Analysis. *Frontiers in Physiology*, 3, 2012.
24. Andrea Lancichinetti and Santo Fortunato. Consensus clustering in complex networks. *Scientific Reports*, 2(1):336, March 2012. Number: 1
25. M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582, June 2006. Number: 23.
26. Lucas G. S. Jeub, Olaf Sporns, and Santo Fortunato. Multiresolution Consensus Clustering in Networks. *Scientific Reports*, 8(1):3259, February 2018.
27. Tarun Kumar, Sankaran Vaidyanathan, Harini Ananthapadmanabhan, Srinivasan Parthasarathy, and Balaraman Ravindran. A New Measure of Modularity in Hypergraphs: Theoretical Insights and Implications for Effective Clustering. In Hocine Cherifi, Sabrina Gaito, José Fernendo Mendes, Esteban Moro, and Luis Mateus Rocha, editors, *Complex Networks and Their Applications VIII*, Studies in Computational Intelligence, pages 286–297, Cham, 2020. Springer International Publishing.
28. Joseph T. Lizier. Measuring the Dynamics of Information Processing on a Local Scale in Time and Space. In Michael Wibral, Raul Vicente, and Joseph T. Lizier, editors, *Directed Information Measures in Neuroscience*, pages 161–193. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. Series Title: Understanding Complex Systems.
29. Joseph T. Lizier, Mikhail Prokopenko, and Albert Y. Zomaya. A Framework for the Local Information Dynamics of Distributed Computation in Complex Systems. In Mikhail Prokopenko, editor, *Guided Self-Organization: Inception*, Emergence, Complexity and Computation, pages 115–158. Springer, Berlin, Heidelberg, 2014.
30. Erik P. Hoel, Larissa Albantakis, and Giulio Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, December 2013. Number: 49.
31. Thomas F. Varley, Vanessa Denny, Olaf Sporns, and Alice Patania. Topological analysis of differential effects of ketamine and propofol anaesthesia on brain dynamics. *Royal Society Open Science*, 8(6):201971, June 2021.