entropy

Article

Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal

Robin A. A. Ince

Institute of Neuroscience and Psychology, University of Glasgow, Glasgow G12 8QB, UK;
robin.ince@glasgow.ac.uk

Received: 2 May 2017; Accepted: 27 June 2017; Published: 29 June 2017

Abstract: The problem of how to properly quantify redundant information is an open question that has been the subject of much recent research. Redundant information refers to information about a target variable S that is common to two or more predictor variables X_i. It can be thought of as quantifying overlapping information content or similarities in the representation of S between the X_i. We present a new measure of redundancy which measures the common change in surprisal shared between variables at the local or pointwise level. We provide a game-theoretic operational definition of unique information, and use this to derive constraints which are used to obtain a maximum entropy distribution. Redundancy is then calculated from this maximum entropy distribution by counting only those local co-information terms which admit an unambiguous interpretation as redundant information. We show how this redundancy measure can be used within the framework of the Partial Information Decomposition (PID) to give an intuitive decomposition of the multivariate mutual information into redundant, unique and synergistic contributions. We compare our new measure to existing approaches over a range of example systems, including continuous Gaussian variables. Matlab code for the measure is provided, including all considered examples.

Keywords: mutual information; redundancy; synergy; pointwise; local; surprisal; partial information decomposition; interaction information; co-information

1. Introduction

Information theory was originally developed as a formal approach to the study of man-made communication systems [1,2]. However, it also provides a comprehensive statistical framework for practical data analysis [3]. For example, mutual information is closely related to the log-likelihood ratio test of independence [4]. Mutual information quantifies the statistical dependence between two (possibly multi-dimensional) variables. When two variables (X and Y) both convey mutual information about a third, S, this indicates that some prediction about the value of S can be made after observing the values of X and Y. In other words, S is represented in some way in X and Y. In many cases, it is interesting to ask how these two representations are related—can the prediction of S be improved by simultaneous observation of X and Y (synergistic representation), or is one alone sufficient to extract all the knowledge about S which they convey together (redundant representation). A principled method to quantify the detailed structure of such representational interactions between multiple variables would be a useful tool for addressing many scientific questions across a range of fields [5–8]. Within the experimental sciences, a practical implementation of such a method would allow analyses that are difficult or impossible with existing statistical methods, but that could provide important insights into the underlying system.

Williams and Beer [6] present an elegant methodology to address this problem, with a non-negative decomposition of multivariate mutual information. Their approach, called the Partial Information Decomposition (PID), considers the mutual information within a set of variables. One variable is considered as a privileged *target* variable, here denoted S, which can be thought of as the independent variable in classical statistics. The PID then considers the mutual information conveyed about this target variable by the remaining *predictor* variables, denoted $\mathcal{X} = \{X_1, X_2, \ldots X_n\}$, which can be thought of as dependent variables. In practice the target variable S may be an experimental stimulus or parameter, while the predictor variables in $\mathcal{X}$ might be recorded neural responses or other experimental outcome measures. However, note that due to the symmetry of mutual information, the framework applies equally when considering a single (dependent) output in response to multiple inputs [7]. Williams and Beer [6] present a mathematical lattice structure to represent the set theoretic intersections of the mutual information of multiple variables [9]. They use this to decompose the mutual information $I(\mathcal{X}; S)$ into terms quantifying the unique, redundant and synergistic information about the independent variable carried by each combination of dependent variables. This gives a complete picture of the representational interactions in the system.

The foundation of the PID is a measure of redundancy between any collection of subsets of $\mathcal{X}$. Intuitively, this should measure the information shared between all the considered variables, or alternatively their common representational overlap. Williams and Beer [6] use a redundancy measure they term I_{min}. However as noted by several authors this measure quantifies the minimum *amount* of information that all variables carry, but does not require that each variable is carrying the *same* information. It can therefore overstate the amount of redundancy in a particular set of variables. Several studies have noted this point and suggested alternative approaches [10–16].

In our view, the additivity of surprisal is the fundamental property of information theory that provides the possibility to meaningfully quantify redundancy, by allowing us to calculate overlapping information content. In the context of the well-known set-theoretical interpretation of information theoretic quantities as measures which quantify the area of sets and which can be visualised with Venn diagrams [9], co-information (often called interaction information) [17–20] is a quantity which measures the intersection of multiple mutual information values (Figure 1). However, as has been frequently noted, co-information conflates synergistic and redundant effects.

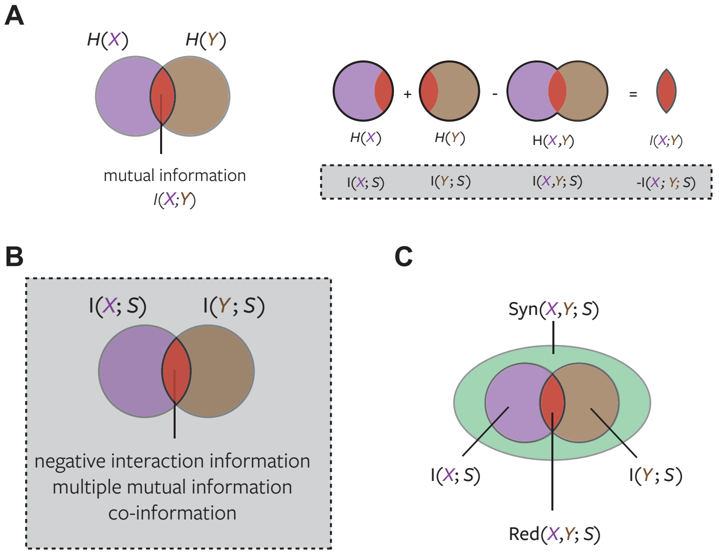

Figure 1. Venn diagrams of mutual information and interaction information. (**A**) Illustration of how mutual information is calculated as the overlap of two entropies; (**B**) The overlapping part of two mutual information values (negative interaction information) can be calculated in the same way—see dashed box in (**A**); (**C**) The full structure of mutual information conveyed by two variables about a third should separate redundant and synergistic regions.

We first review co-information and the PID before presenting I_{ccs}, a new measure of redundancy based on quantifying the common change in surprisal between variables at the local or pointwise level [21–25]. We provide a game-theoretic operational motivation for a set of constraints over which we calculate the maximum entropy distribution. This game-theoretic operational argument extends the decision theoretic operational argument of [12] but arrives at different conclusions about the fundamental nature of unique information. We demonstrate the PID based on this new measure with several examples that have been previously considered in the literature. Finally, we apply the new measure to continuous Gaussian variables [26].

2. Interaction Information (Co-Information)

2.1. Definitions

The foundational quantity of information theory is *entropy*, which is a measure of the variability or uncertainty of a probability distribution. The entropy of a discrete random variable X, with probability mass function $P(X)$ is defined as:

$$H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)} \tag{1}$$

This is the expectation over X of $h(x) = -\log_2 p(x)$, which is called the *surprisal* of a particular value x. If a value x has a low probability, it has high surprisal and vice versa. Many information theoretic quantities are similarly expressed as an expectation—in such cases, the specific values of the function over which the expectation is taken are called *pointwise* or *local* values [21–25]. We denote these local values with a lower case symbol. Following [7] we denote probability distributions with a capital latter, e.g., $P(X_1, X_2)$, but denote values of specific realisations, i.e., $P(X_1 = x_1, X_2 = x_2)$ with lower case shorthand $p(x_1, x_2)$.

Figure 1A shows a Venn diagram representing the entropy of two variables X and Y. One way to derive mutual information $I(X;Y)$ is as the intersection of the two entropies. This intersection can be calculated directly by summing the individual entropies (which counts the overlapping region twice) and subtracting the joint entropy (which counts the overlapping region once). This matches one of the standard forms of the definition of mutual information:

$$\begin{align} I(X;Y) &= H(X) + H(Y) - H(X,Y) \tag{2} \\ &= \sum_{x,y} p(x,y) \left[\log_2 \frac{1}{p(y)} - \log_2 \frac{1}{p(y|x)} \right] \tag{3} \end{align}$$

Here $p(y|x)$ denotes the conditional probability of observing $Y = y$, given that $X = x$ has been observed: $p(y|x) = {}^{p(y,x)}/_{p(x)}$. Mutual information is the expectation of $i(x;y) = h(y) - h(y|x) = \log_2 \frac{p(y|x)}{p(y)}$, the difference in surprisal of value y when value x is observed. To emphasise this point we use a notation which makes explicit the fact that pointwise information measures a change in surprisal

$$\begin{align} i(x;y) &= \Delta_y h(x) = h(x) - h(x|y) \tag{4} \\ &= \Delta_x h(y) = h(y) - h(y|x) \tag{5} \end{align}$$

Mutual information is non-negative, symmetric and equals zero if and only if the two variables are statistically independent (that is, $p(x,y) = p(x)p(y) \forall x \in X, y \in Y$) [2].

A similar approach can be taken when considering mutual information about a target variable S that is carried by two predictor variables X and Y (Figure 1B). Again the overlapping region can be calculated directly by summing the two separate mutual information values and subtracting the joint information. However, in this case the resulting quantity can be negative. Positive values of the intersection represent a net redundant representation: X and Y share the same information about S.

Negative values represent a net synergistic representation: X and Y provide more information about S together than they do individually.

In fact, this quantity was first defined as the negative of the intersection described above, and termed *interaction information* [17]:

$$\begin{aligned} I(X;Y;S) &= I(X,Y;S) - I(X;S) - I(Y;S) \\ &= I(S;X|Y) - I(S;X) \\ &= I(S;Y|X) - I(S;Y) \\ &= I(X;Y|S) - I(X;Y) \end{aligned} \tag{6}$$

The alternative equivalent formulations illustrate how the interaction information is symmetric in the three variables, and also represents for example, the information between S and X which is gained (synergy) or lost (redundancy) when Y is fixed (conditioned out).

This quantity has also been termed *multiple mutual information* [27], *co-information* [19], *higher-order mutual information* [20] and *synergy* [28–31]. Multiple mutual information and co-information use a different sign convention from interaction information. For odd numbers of variables (e.g., three X_1, X_2, S) co-information has the opposite sign to interaction information; positive values indicate net redundant overlap.

As for mutual information and conditional mutual information, the interaction information as defined above is an expectation over the joint probability distribution. Expanding the definitions of mutual information in Equation (6) gives:

$$I(X;Y;S) = \sum_{x,y,s} p(x,y,s) \log_2 \frac{p(x,y,s)p(x)p(y)p(s)}{p(x,y)p(x,s),p(y,s)} \tag{7}$$

$$I(X;Y;S) = \sum_{x,y,s} p(x,y,s) \left[\log_2 \frac{p(s|x,y)}{p(s)} - \log_2 \frac{p(s|x)}{p(s)} - \log_2 \frac{p(s|y)}{p(s)} \right] \tag{8}$$

As before we can consider the local or pointwise function

$$i(x;y;s) = \Delta_s h(x,y) - \Delta_s h(x) - \Delta_s h(y) \tag{9}$$

The negation of this value measures the overlap in the change of surprisal about s between values x and y (Figure 1A).

It can be seen directly from the definitions above that in the three variable case the interaction information is bounded:

$$\begin{aligned} I(X;Y;S) &\geq -\min\left[I(S;X), I(S;Y), I(X;Y)\right] \\ I(X;Y;S) &\leq \min\left[I(S;X|Y), I(S;Y|X), I(X;Y|S)\right] \end{aligned} \tag{10}$$

We have introduced interaction information for three variables, from a perspective where one variable is privileged (independent variable) and we study interactions in the representation of that variable by the other two. However, as noted interaction information is symmetric in the arguments, and so we get the same result whichever variable is chosen to provide the analysed information content.

Interaction information is defined similarly for larger numbers of variables. For example, with four variables, maintaining the perspective of one variable being privileged, the 3-way Venn diagram intersection of the mutual information terms again motivates the definition of interaction information:

$$\begin{aligned} I(W;X;Y;S) = &- I(W;S) - I(X;S) - I(Y;S) \\ &+ I(W,X;S) + I(W,Y;S) + I(Y,X;S) \\ &- I(W,X,Y;S) \end{aligned} \tag{11}$$

In the n-dimensional case the general expression for interaction information on a variable set $\mathcal{V} = \{\mathcal{X}, S\}$ where $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ is:

$$I(\mathcal{V}) = -\sum_{\mathcal{T} \subseteq \mathcal{X}} (-1)^{|\mathcal{T}|} I(\mathcal{T}; S) \tag{12}$$

which is an alternating sum over all subsets $\mathcal{T} \subseteq \mathcal{X}$, where each $\mathcal{T}$ contains $|\mathcal{T}|$ elements of $\mathcal{X}$. The same expression applies at the local level, replacing I with the pointwise i. Dropping the privileged target S an equivalent formulation of interaction information on a set of n-variables $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ in terms of entropy is given by [18,32]:

$$I(\mathcal{X}) = -\sum_{\mathcal{T} \subseteq \mathcal{X}} (-1)^{|\mathcal{X}|-|\mathcal{T}|} H(\mathcal{T}) \tag{13}$$

2.2. Interpretation

We consider as above a three variable system with a target variable S and two predictor variables X, Y, with both X and Y conveying information about S. The concept of redundancy is related to whether the information conveyed by X and that conveyed by Y is *the same* or *different*. Within a decoding (supervised classification) approach, the relationship between the variables is determined from predictive performance within a cross-validation framework [33,34]. If the performance when decoding X and Y together is the same as the performance when considering e.g., X alone, this indicates that the information in Y is completely redundant with that in X; adding observation of Y has no predictive benefit for an observer. In practice redundancy may not be complete as in this example; some part of the information in X and Y might be shared, while both variables also convey unique information not available in the other.

The concept of synergy is related to whether X and Y convey more information when observed together than they do when observed independently. Within the decoding framework this means higher performance is obtained by a decoder which predicts on a joint model of simultaneous X and Y observations, versus a decoder which combines independent predictions obtained from X and Y individually. The predictive decoding framework provides a useful intuition for the concepts, but has problems quantifying redundancy and synergy in a meaningful way because of the difficulty of quantitatively relating performance metrics (percent correct, area under ROC, etc.) between different sets of variables—i.e., X, Y and the joint variable (X, Y).

The first definition (Equation (6)) shows that interaction information is the natural information theoretic approach to this problem: it contrasts the information available in the joint response to the information available in each individual response (and similarly obtains the intersection of the multivariate mutual information in higher order cases). A negative value of interaction information quantifies the redundant overlap of Figure 1B, positive values indicate a net synergistic effect between the two variables. However, there is a major issue which complicates this interpretation: interaction information conflates synergy and redundancy in a single quantity (Figure 1B) and so does not provide a mechanism for separating synergistic and redundant information (Figure 1C) [6]. This problem arises for two reasons. First, local terms $i(x; y; s)$ can be positive for some values of x, y, s and negative for others. These opposite effects can then cancel in the overall expectation. Second, as we will see, the computation of interaction information can include terms which do not have a clear interpretation in terms of synergy or redundancy.

3. The Partial Information Decomposition

In order to address the problem of interaction information conflating synergistic and redundant effects, Williams and Beer [6] proposed a decomposition of mutual information conveyed by a set of predictor variables $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$, about a target variable S. They reduce the total multivariate mutual information, $I(\mathcal{X}; S)$, into a number of non-negative atoms representing the unique, redundant and synergistic information between all subsets of $\mathcal{X}$: in the two-variable case this corresponds to the four regions of Figure 1C. To do this they consider all subsets of $\mathcal{X}$, denoted $\mathbf{A_i}$, and termed *sources*. They show that the redundancy structure of the multivariate information is determined by the "collection of all sets of sources such that no source is a superset of any other"—formally the set of anti-chains on the lattice formed from the power set of $\mathcal{X}$ under set inclusion, denoted $\mathcal{A}(\mathcal{X})$. Together with a natural ordering, this defines a redundancy lattice [35]. Each node of the lattice represents a partial information atom, the value of which is given by a partial information (PI) function. Note there is a direct correspondence between the lattice structure and a Venn diagram representing multiple mutual information values. Each node on a lattice corresponds to a particular intersecting region in the Venn diagram. For two variables there are only four terms, but the advantage of the lattice representation becomes clearer for higher number of variables. The lattice view is much easier to interpret when there are a large number of intersecting regions that are hard to visualise in a Venn diagram. Figure 2 shows the structure of this lattice for $n = 2, 3$. The PI value for each node, denoted I_∂, can be determined via a recursive relationship (Möbius inverse) over the redundancy values of the lattice:

$$I_\partial(S;\alpha) = I_\cap(S;\alpha) - \sum_{\beta \prec \alpha} I_\partial(S;\beta) \tag{14}$$

where $\alpha \in \mathcal{A}(\mathcal{X})$ is a set of sources (each a set of input variables X_i) defining the node in question.

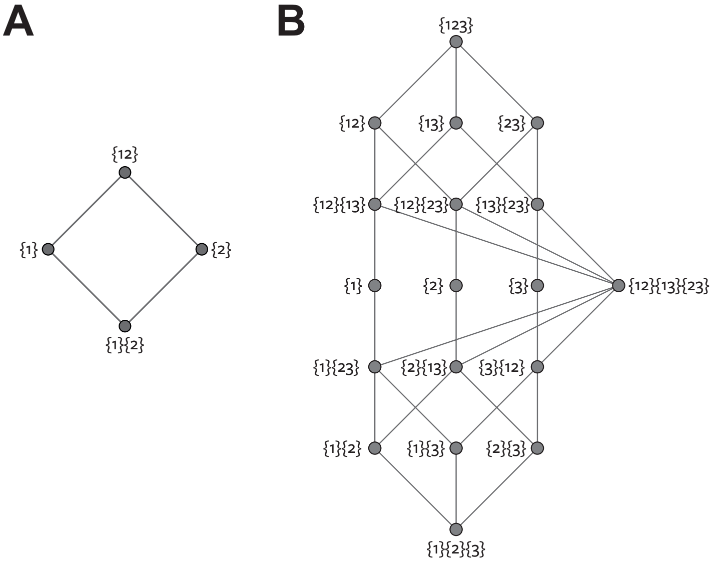

Figure 2. Redundancy lattice for (**A**) two variables; (**B**) three variables. Modified from [6].

The redundancy value of each node of the lattice, $I_\cap$, measures the total amount of redundant information shared between the sources included in that node. For example, $I_\cap(S; \{X_1\}\{X_2\})$ quantifies the redundant information content about S that is common to both X_1 and X_2. The partial information function, I_∂, measures the unique information contributed by only that node (redundant, synergistic or unique information within subsets of variables).

For the two variable case, if the redundancy function used for a set of sources is denoted $I_\cap\left(S; \mathbf{A_1}, \ldots, \mathbf{A_k}\right)$ and following the notation in [6], the nodes of the lattice, their redundancy and their partial information values are given in Table 1.

Table 1. Full Partial Information Decomposition (PID) in the two-variable case. The four terms here correspond to the four regions in Figure 1C.

Node Label	Redundancy Function	Partial Information	Represented Atom
{12}	$I_\cap(S;\{X_1,X_2\})$	$I_\cap(S;\{X_1,X_2\})$ $-\,I_\cap(S;\{X_1\}) - I_\cap(S;\{X_2\})$ $+\,I_\cap(S;\{X_1\}\{X_2\})$	unique information in X_1 and X_2 together (synergy)
{1}	$I_\cap(S;\{X_1\})$	$I_\cap(S;\{X_1\})$ $-\,I_\cap(S;\{X_1\}\{X_2\})$	unique information in X_1 only
{2}	$I_\cap(S;\{X_2\})$	$I_\cap(S;\{X_2\})$ $-\,I_\cap(S;\{X_1\}\{X_2\})$	unique information in X_2 only
{1}{2}	$I_\cap(S;\{X_1\}\{X_2\})$	$I_\cap(S;\{X_1\}\{X_2\})$	redundant information between X_1 and X_2

Note that we have not yet specified a redundancy function. A number of axioms have been proposed for any candidate redundancy measure [6,11]:

Symmetry:

$$I_\cap\left(S;\mathbf{A_1},\ldots,\mathbf{A_k}\right) \text{ is symmetric with respect to the } \mathbf{A_i}\text{'s.} \tag{15}$$

Self Redundancy:

$$I_\cap\left(S;\mathbf{A}\right) = I(S;\mathbf{A}) \tag{16}$$

Subset Equality:

$$I_\cap\left(S;\mathbf{A_1},\ldots,\mathbf{A_{k-1}},\mathbf{A_k}\right) = I_\cap\left(S;\mathbf{A_1},\ldots,\mathbf{A_{k-1}}\right) \text{ if } \mathbf{A_{k-1}} \subseteq \mathbf{A_k} \tag{17}$$

Monotonicity:

$$I_\cap\left(S;\mathbf{A_1},\ldots,\mathbf{A_{k-1}},\mathbf{A_k}\right) \leq I_\cap\left(S;\mathbf{A_1},\ldots,\mathbf{A_{k-1}}\right) \tag{18}$$

Note that previous presentations of these axioms have included subset equality as part of the monotonicity axiom; we separate them here for reasons that will become clear later. Subset equality allows the full power set of all combinations of sources to be reduced to only the anti-chains under set inclusion (the redundancy lattice). Self redundancy ensures that the top node of the redundancy lattice, which contains a single source $\mathbf{A} = \mathcal{X}$, is equal to the full multivariate mutual information and therefore the lattice structure can be used to decompose that quantity. Monotonicity ensures redundant information is increasing with the height of the lattice, and has been considered an important requirement that redundant information should satisfy.

Other authors have also proposed further properties and axioms for measures of redundancy [13,14]. In particular, Reference [11] propose an additional axiom regarding the redundancy between two sources about a variable constructed as a copy of those sources:

Identity Property (Harder et al.):

$$I_\cap\left(\left[\mathbf{A_1},\mathbf{A_2}\right];\mathbf{A_1},\mathbf{A_2}\right) = I(\mathbf{A_1};\mathbf{A_2}) \tag{19}$$

In this manuscript we focus on redundant and synergistic mutual information. However, the concepts of redundancy and synergy can also be applied directly to entropy [36]. Redundant entropy is variation that is shared between two (or more) variables, synergistic entropy is additional uncertainty that arises when the variables are considered together, over and above what would be obtained if they were statistically independent. Note that since the global joint entropy quantity is maximised when the two variables are independent, redundant entropy is always greater than synergistic entropy [36]. However, local synergistic entropy can still occur: consider negative local information terms, which by definition quantify a synergistic local contribution to the joint entropy sum since $h(x, y) > h(x) + h(y)$. A crucial insight that results from this point of view is that mutual information itself quantifies both redundant and synergistic entropy effects—it is the difference between redundant and synergistic entropy across the two inputs [36]. With H_∂ denoting redundant or synergistic partial entropy analogous to partial information we have:

$$I(\mathbf{A_1}; \mathbf{A_2}) = H_\partial(\{\mathbf{A_1}\}\{\mathbf{A_2}\}) - H_\partial(\{\mathbf{A_1}, \mathbf{A_2}\}) \tag{20}$$

This is particularly relevant for the definition of the identity axiom. We argue that the previously unrecognised contribution of synergistic entropy to mutual information (pointwise negative terms in the mutual information expectation sum) should not be included in an information redundancy measure.

Note that any information redundancy function can induce an entropy redundancy function by considering the information redundancy with the copy of the inputs. For example, for the bivariate case we can define:

$$H_\cap(\{\mathbf{A_1}\}\{\mathbf{A_2}\}) = I_\cap\left([\mathbf{A_1}, \mathbf{A_2}]\,; \mathbf{A_1}, \mathbf{A_2}\right) \tag{21}$$

So any information redundancy measure that satisfies the identity property [12] cannot measure synergistic entropy [36], since for the induced entropy redundancy measure $H_\cap(\{\mathbf{A_1}\}\{\mathbf{A_2}\}) = I(\mathbf{A_1}; \mathbf{A_2})$ so from Equation (20) $H_\partial(\{\mathbf{A_1}\mathbf{A_2}\}) = 0$. To address this without requiring introducing in detail the partial entropy decomposition [36], we propose a modified version of the identity axiom, which still addresses the two-bit copy problem but avoids the problem of including synergistic mutual information contributions in the redundancy measure. When $I(\mathbf{A_1}; \mathbf{A_2}) = 0$ there are no synergistic entropy effects because $i(a_1, a_2) = 0\ \forall a_1, a_2$ so there are no misinformation terms and no synergistic entropy between the two inputs.

Independent Identity Property:

$$I(\mathbf{A_1}; \mathbf{A_2}) = 0 \implies I_\cap\left([\mathbf{A_1}, \mathbf{A_2}]\,; \mathbf{A_1}, \mathbf{A_2}\right) = 0 \tag{22}$$

Please note that while this section primarily reviews existing work on the partial information decomposition, two novel contributions here are the explicit consideration of subset equality separate to monotonicity, and the definition of the independent identity property.

3.1. *An Example PID:* RdnUnqXor

Before considering specific measures of redundant information that have been proposed for use with the PID, we first illustrate the relationship between the redundancy and the partial information lattice values with an example. We consider a system called RdnUnqXor [10]. The structure of this system is shown in Figure 3A [37]. It consists of two three bit predictors, X_1 and X_2, and a four bit target S. This example is noteworthy, because an intuitive PID is obvious from the definition of the system, and it includes by construction 1 bit of each type of information decomposable with the PID.

All three variables share a bit (labelled b in Figure 3A). This means there should be 1 bit of redundant information. Bit b is shared between each predictor and the target so forms part of $I(X_i; S)$, and is also shared between the predictors, therefore it is shared or redundant information. All variables have one bit that is distributed according to a Xor configuration across the three variables (labelled a). This provides 1 bit of synergy within the system, because the value of bit a of S can only be predicted when X_1 and X_2 are observed together simultaneously [10]. Bits c and d are shared between S and each of X_1 and X_2 individually. So each of these contributes to $I(X_i; S)$, but as unique information.

We illustrate the calculation of the PID for this system (Figure 3B,C, Table 2). From the self-redundancy axiom, the three single-source terms can all be calculated directly from the classical mutual information values. The single predictors each have 2 bits of mutual information (the two bits shared with S). Both predictors together have four bits of mutual information with S, since the values of all four bits of S are all fully determined when both X_1 and X_2 are observed. Since by construction there is 1 bit shared redundantly between the predictors, we claim $I_\cap(S;\{1\}\{2\}) = 1$ bit and we have all the redundancy values on the lattice. Then from the summation procedure illustrated in Table 1 we can calculate the partial information values. For example, $I_\partial(S;\{1\}) = 2 - 1 = 1$, and $I_\partial(S;\{12\}) = 4 - 1 - 1 - 1 = 1$.

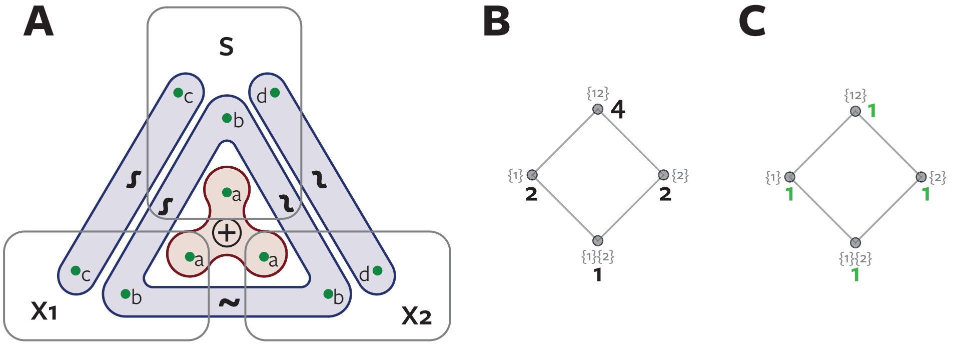

Figure 3. Partial Information Decomposition for RdnUnqXor (**A**) The structure of the RdnUnqXor system borrowing the graphical representation from [37]. S is a variable containing 4 bits (labelled a, b, c, d). X_1 and X_2 each contain 3 bits. $\sim$ indicates bits which are coupled (distributed identically) and $\oplus$ indicates the enclosed variables form the Xor relation; (**B**) Redundant information values on the lattice (black); (**C**) Partial information values on the lattice (green).

Table 2. PID for RdnUnqXor (Figure 3).

Node	$I_\cap$	I_∂
{1}{2}	1	1
{1}	2	1
{2}	2	1
{12}	4	1

3.2. Measuring Redundancy With Minimal Specific Information: I_{min}

The redundancy measure proposed by Williams and Beer [6] is denoted $I_{\min}$ and derived as the average (over values s of S) minimum specific information [38,39] over the considered input sources. The information provided by a source $\mathbf{A}$ (as above a subset of dependent variables X_i) can be written:

$$I(S;\mathbf{A}) = \sum_s p(s) I(S = s;\mathbf{A}) \tag{23}$$

where $I(S = s;\mathbf{A})$ is the *specific information*:

$$I(S = s;\mathbf{A}) = \sum_{\mathbf{a}} p(\mathbf{a}|s) \left[\log_2 \frac{1}{p(s)} - \log_2 \frac{1}{p(s|\mathbf{a})} \right] \tag{24}$$

which quantifies the average reduction in surprisal of s given knowledge of $\mathbf{A}$. This splits the overall mutual information into the reduction in uncertainty about each individual target value. $I_{\min}$ is then defined as:

$$I_{\min}(S;\mathbf{A_1},\ldots,\mathbf{A_k}) = \sum_s p(s) \min_{\mathbf{A_i}} I(S = s;\mathbf{A_i}) \tag{25}$$

This quantity is the expectation (over S) of the minimum amount of information about each specific target value s conveyed by any considered source. $I_{\min}$ is non-negative and satisfies the axioms of symmetry, self redundancy and monotonicity, but not the identity property (neither Harder et al. or independent forms). The crucial conceptual problem with $I_{\min}$ is that it indicates the variables share a common *amount* of information, but not that they actually share the *same* information content [5,10,11].

The most direct example of this is the "two-bit copy problem", which motivated the identity axiom [5,10,11]. We consider two independent uniform binary variables X_1 and X_2 and define S as a direct copy of these two variables $S = (X_1, X_2)$. In this case $I_{\min}(S;\{1\}\{2\}) = 1$ bit; for every s both X_1 and X_2 each provide 1 bit of specific information. However, both variables give different information about each value of s: X_1 specifies the first component, X_2 the second. Since X_1 and X_2 are independent by construction there should be no overlap. This illustrates that $I_{\min}$ can overestimate redundancy with respect to an intuitive notion of overlapping information content.

3.3. Measuring Redundancy With Maximised Co-Information: I_{broja}

A number of alternative redundancy measures have been proposed for use with the PID in order to address the problems with $I_{\min}$ (reviewed by [26]). Two groups have proposed an equivalent approach, based on the idea that redundancy should arise only from the marginal distributions $P(X_1, S)$ and $P(X_2, S)$ ([12], their Assumption *) and that synergy should arise from structure not present in those two marginals, but only in the full joint distribution $P(X_1, X_2, S)$. Please note that we follow their terminology and refer to this concept as Assumption * throughout. Griffith and Koch [10] frame this view as a minimisation problem for the multivariate information $I(S; X_1, X_2)$ over the class of distributions which preserve the individual source-target marginal distributions. Bertschinger et al. [12] seek to minimise $I(S; X_1|X_2)$ over the same class of distributions, but as noted both approaches result in the same PID. In both cases the redundancy, $I_\cap(S;\{X_1\}\{X_2\})$, is obtained as the maximum of the co-information (negative interaction information) over all distributions that preserve the source-target marginals:

$$I_{\text{max-nii}}(S;\{X_1\}\{X_2\}) = \max_{Q\in\Delta_P} -I_Q(S;X_1;X_2) \tag{26}$$

$$\Delta_P = \{Q \in \Delta : Q(X_1,S) = P(X_1,S), Q(X_2,S) = P(X_2,S)\} \tag{27}$$

We briefly highlight here a number of conceptual problems with this approach. First, this measure satisfies the Harder et al. identity property (Equation (19)) [11,12] and is therefore incompatible with the notion of synergistic entropy [36]. Second, this measure optimises co-information, a quantity which conflates synergy and redundancy [6]. Given ([12], Assumption *) which states that unique and redundant information are constant on the optimisation space, this is equivalent to minimizing synergy [7].

$$I_{\text{max-nii}}(S;\{X_1\}\{X_2\}) = I_{\text{red}}(S;\{X_1\}\{X_2\}) - I_{\text{syn-min}}(S;\{X_1\}\{X_2\}) \tag{28}$$

where $I_{\text{syn-min}}(S;\{X_1\}\{X_2\})$ is the smallest possible synergy given the target-predictor marginal constraints, but is not necessarily zero. Therefore, the measure provides a bound on redundancy (under Assumption * [12]) but cannot measure the true value. Third, Bertschinger et al. [12] motivate the constraints for the optimisation from an operational definition of unique information based on decision theory. It is this argument which suggests that the unique information is constant on the optimisation space Δ_P, and which motivates a foundational axiom for the measure that equal target-predictor marginal distributions imply zero unique information. However, we do not agree that unique information is invariant to the predictor-predictor marginal distributions, or necessarily equals zero when target-predictor marginals are equal. We revisit the operational definition in Section 4.3 by considering a game theoretic extension which provides a different perspective. We use this to provide a counter-example that proves the decision theoretic argument is not a necessary condition for the existence of unique information, and therefore the I_{broja} procedure is invalid since redundancy is not fixed on Δ_P. We also demonstrate with several examples (Section 5) how the I_{broja} optimisation results in coupled predictor variables, suggesting the co-information optimisation is indeed maximising the source redundancy between them.

3.4. Other Redundancy Measures

Harder et al. [11] define a redundancy measure based on a geometric projection argument, which involves an optimisation over a scalar parameter λ, and is defined only for two sources, so can be used only for systems with two predictor variables. Griffith et al. [13] suggest an alternative measure motivated by zero-error information, which again formulates an optimisation problem (here maximisation of mutual information) over a family of distributions (here distributions Q which are a function of each predictor so that $H(Q|X_i) = 0$). Griffith and Ho [16] extend this approach by modifying the optimisation constraint to be $H(Q|X_i) = H(Q|X_i, Y)$.

4. Measuring Redundancy With Pointwise Common Change in Surprisal: I_{ccs}

We derive here from first principles a measure that we believe encapsulates the intuitive meaning of redundancy between sets of variables. We argue that the crucial feature which allows us to directly relate information content between sources is the additivity of surprisal. Since mutual information measures the expected change in pointwise surprisal of s when x is known, we propose measuring redundancy as the expected pointwise change in surprisal of s which is common to x and y. We term this *common change in surprisal* and denote the resulting measure $I_{\text{ccs}}(S;\alpha)$.

4.1. Derivation

As for entropy and mutual information we can consider a Venn diagram (Figure 1) for the change in surprisal of a specific value s for specific values x and y and calculate the overlap directly using local co-information (negative local interaction information). However, as noted before the interaction information can confuse synergistic and redundant effects, even at the pointwise level. Recall that mutual information $I(S;X)$ is the expectation of a local function which measures the pointwise change in surprisal $i(s;x) = \Delta_s h(x)$ of value s when value x is observed. Although mutual information itself is always non-negative, the pointwise function can take both positive and negative values. Positive values correspond to a reduction in the surprisal of s when x is observed, negative values to an increase in surprisal. Negative local information values are sometimes referred to as *misinformation* [23] and can be interpreted as representing synergistic entropy between S and X [36]. Mutual information is then the expectation of both positive (information) terms and negative (misinformation) terms. Table 3 shows how the possibility of local misinformation terms complicates pointwise interpretation of the local negative interaction information (co-information).

Note that the fourth column represents the local co-information which quantifies the set-theoretic overlap of the two univariate local information values. By considering the signs of all four terms, the two univariate local informations, the local joint information and their overlap, we can determine terms which correspond to redundancy and terms which correspond to synergy. We make an assumption that a decrease in surprisal of s (positive local information term) is a fundamentally different event to an increase in surprisal of s (negative local information). Therefore, we can only interpret the local co-information as a set-theoretic overlap in the case where all three local information terms have the same sign. If the joint information has a different sign to the individual informations (rows 5 and 6) the two variables together represent a fundamentally different change in surprisal than either do alone. While a full interpretation of what these terms might represent is difficult, we argue it is clear they cannot represent a common change in surprisal. Similarly, if the two univariate local informations have opposite sign, they cannot have any common overlap.

Table 3. Different interpretations of local interaction information terms. ? indicates that combination of terms does not admit a clear interpretation in terms of redundancy or synergy.

$\Delta_s h(x)$	$\Delta_s h(y)$	$\Delta_s h(x,y)$	$-i(x;y;s)$	**Interpretation**
$+$	$+$	$+$	$+$	redundant information
$+$	$+$	$+$	$-$	synergistic information
$-$	$-$	$-$	$-$	redundant misinformation
$-$	$-$	$-$	$+$	synergistic misinformation
$+$	$+$	$-$	$\ldots$	?
$-$	$-$	$+$	$\ldots$	?
$+/-$	$-/+$	$\ldots$	$\ldots$	?

The table shows that interaction information combines redundant information with synergistic misinformation, and redundant misinformation with synergistic information. As discussed, it also includes terms which do not admit a clear interpretation. We argue that a principled measure of redundancy should consider only redundant information and redundant misinformation. We therefore consider the pointwise negative interaction information (overlap in surprisal), but only for symbols corresponding to the first and third rows of Table 3. That is, terms where the sign of the change in surprisal for all the considered sources is equal, and equal also to the sign of overlap (measured with local co-information). In this way, we count the contributions to the overall mutual information (both positive and negative) which are genuinely shared between the input sources, while ignoring other (synergistic and ambiguous) interaction effects. We assert that conceptually this is exactly what a redundancy function should measure.

We denote the local co-information (negative interaction information if n is odd) with respect to a joint distribution Q as $c_q(a_1,\ldots,a_n)$, which is defined as [20]:

$$c_q(a_1,\ldots,a_n) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1<\cdots<i_k} h_q\left(a_{i_1},\ldots,a_{i_k}\right) \tag{29}$$

where $h_q(a_1,\ldots,a_n) = -\log q(a_1,\ldots,a_n)$ is pointwise entropy (surprisal). Then we define I_{ccs}, the common change in surprisal, as:

Definition 1.

$$\begin{aligned} I_{\text{ccs}}(S;A_1,\ldots,A_n) &= \sum_{a_1,\ldots,a_n} \tilde{p}(a_1,\ldots,a_n)\Delta_s h^{\text{com}}(a_1,\ldots,a_n) \\ \Delta_s h^{\text{com}}(a_1,\ldots,a_n) &= \begin{cases} c_{\tilde{p}}(a_1,\ldots,a_n,s) & \text{if} \quad \operatorname{sgn}\Delta_s h(a_1) = \ldots = \operatorname{sgn}\Delta_s h(a_n) \\ & \qquad = \operatorname{sgn}\Delta_s h(a_1,\ldots,a_n) = \operatorname{sgn} c(a_1,\ldots,a_n,s) \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{30}$$

where $\Delta_s h^{\text{com}}(a_1,\ldots,a_n)$ represents the common change in surprisal (which can be positive or negative) between the input source values, and $\tilde{P}$ is a joint distribution obtained from the observed joint distribution P (see below). I_{ccs} measures overlapping information content with co-information, by separating contributions which correspond unambiguously to redundant mutual information at the pointwise level, and taking the expectation over these local redundancy values.

Unlike I_{min} which considered each input source individually, the pointwise overlap computed with local co-information requires a joint distribution over the input sources, $\tilde{P}$ in order to obtain the local surprisal values $h_{\tilde{p}}(a_1,\ldots,a_n,s)$. We use the maximum entropy distribution subject to the constraints of equal bivariate source-target marginals, together with the equality of the n-variate joint target marginal distribution:

Definition 2.

$$\begin{aligned} \hat{P}(A_1, \ldots, A_n, S) &= \underset{Q \in \Delta_P}{\arg\max} \sum_{a_1,\ldots,a_n,s} -q(a_1, \ldots, a_n, s) \log q(a_1, \ldots, a_n, s) \\ \Delta_P &= \left\{ Q \in \Delta : \begin{array}{l} Q(A_i, S) = P(A_i, S) \text{ for } i = 1, \ldots, n \\ Q(A_1, \ldots, A_n) = P(A_1, \ldots, A_n) \end{array} \right\} \end{aligned} \tag{31}$$

where $P(A_1, \ldots, A_n, S)$ is the probability distribution defining the system under study and here Δ is the set of all possible joint distributions on $A_1, \ldots, A_n, S$. We develop the motivation for the constraints in Section 4.3.1, and for using the distribution with maximum entropy subject to these constraints in in Section 4.3.2.

In a previous version of this manuscript we used constraints obtained from the decision theoretic operational definition of unique information [12]. We used the maximum entropy distribution subject to the constraints of pairwise target-predictor marginal equality:

Definition 3.

$$\begin{aligned} \hat{P}_{\text{ind}}(A_1, \ldots, A_n, S) &= \underset{Q \in \Delta_P}{\arg\max} \sum_{a_1,\ldots,a_n,s} -q(a_1, \ldots, a_n, s) \log q(a_1, \ldots, a_n, s) \\ \Delta_P &= \left\{ Q \in \Delta : \; Q(A_i, S) = P(A_i, S) \text{ for } i = 1, \ldots, n \right\} \end{aligned} \tag{32}$$

This illustrates I_{ccs} can be defined in a way compatible with either operational perspective, depending on whether it is calculated using $\hat{P}$ or $\hat{P}_{\text{ind}}$. We suggest that if a reader favours the decision theoretic definition of unique information [12] over the new game-theoretic definition proposed here (Section 4.3.1) I_{ccs} can be defined in a way consistent with that, and still provides advantages over I_{broja}, which maximises co-information without separating redundant from synergistic contributions (Sections 3.3 and 4.3.2). We include Definition 3 here for continuity with the earlier version of this manuscript, but note that for all the examples considered here we use $\hat{P}$, following the game theoretic operational definition of unique information (Section 4.3.1).

Note that the definition of I_{min} in terms of minimum specific information [39] (Equation (25)) suggests as a possible extension the use of a form of *specific co-information*. In order to separate redundant from synergistic components this should be thresholded with zero to only count positive (redundant) contributions. This can be defined both in terms of target-specific co-information following I_{min} (for clarity these definitions are shown only for two variable inputs):

$$I_{\text{target specific coI}}(S; \mathbf{A_1}, \mathbf{A_2}) = \sum_s p(s) \max\left[I(S = s; \mathbf{A_1}) + I(S = s; \mathbf{A_2}) - I(S = s; \mathbf{A_1}, \mathbf{A_2}), 0 \right] \tag{33}$$

or alternatively in terms of source-specific co-information:

$$I_{\text{source specific coI}}(S; \mathbf{A_1}, \mathbf{A_2}) = \sum_{a_1,a_2} p(a_1, a_2) \max\left[I(S; \mathbf{A_1} = a_1) + I(S; \mathbf{A_2} = a_2) \right. \tag{34}$$

$$\left. - \, I(S; \mathbf{A_1} = a_1, \mathbf{A_2} = a_2), 0 \right] \tag{35}$$

I_{ccs} can be seen as a fully local approach within this family of measures. The first key ingredient of this family is to exploit the additivity of surprisal and hence use the co-information to quantify the overlapping information content (Figure 1); the second ingredient is to break down in some way the expectation summation in the calculation of co-information, to separate redundant and synergistic effects that are otherwise conflated. We argue the fully local view of I_{ccs} is required to fully separate redundant from synergistic effects. In either specific co-information calculation, when summing the contributions within the expectation over the non-specific variable any combination of terms listed in Table 3 could occur. Therefore, these specific co-information values could still conflate redundant information with synergistic misinformation.

4.2. Calculating I_{ccs}

We provide here worked examples of calculating I_{ccs} for two simple example systems. The simplest example of redundancy is when the system consists of a single coupled bit [10] (Example RDN), defined by the following distribution $P(X_1, X_2, S)$:

$$p(0,0,0) = p(1,1,1) = 0.5 \tag{36}$$

In this example $\hat{P} = P$; the maximum entropy optimisation results in the original distribution. Table 4 shows the pointwise terms of the co-information calculation. In this system for both possible configurations the change in surprisal from each predictor is 1 bit and overlaps completely. The signs of all changes in surprisal and the local co-information are positive, indicating that both these events correspond to redundant local information. In this case I_{ccs} is equal to the co-information.

Table 4. Pointwise values from $I_{\text{ccs}}(S; \{1\}\{2\})$ for RDN.

(x_1, x_2, s)	$\Delta_s h(x_1)$	$\Delta_s h(x_2)$	$\Delta_s h(x_1, x_2)$	$c(x_1; x_2; s)$	$\Delta_s h^{\text{com}}(x_1, x_2)$
$(0,0,0)$	1	1	1	1	1
$(1,1,1)$	1	1	1	1	1

The second example we consider is binary addition (see also Section 5.2.2), $S = X_1 + X_2$, with distribution $P(X_1, X_2, S)$ given by

$$p(0,0,0) = p(0,1,1) = p(1,0,1) = p(1,1,2) = 1/4 \tag{37}$$

In this example, again $\hat{P} = P$. The pointwise terms are shown in Table 5. For the events with $x_1 = x_2$, both predictors provide 1 bit local change in surprisal of s, but they do so independently since the change in surprisal when observing both together is 2 bits. Therefore, the local co-information is 0; there is no overlap. For the terms where $x_1 \neq x_2$, neither predictor alone provides any local information about s. However, together they provide a 1 bit change in surprisal. This is therefore a purely synergistic contribution, providing -1 bits of local co-information. However, since this is synergistic, it is not included in $\Delta_s h^{\text{com}}$. $I_{\text{ccs}}(S; \{1\}\{2\}) = 0$, although the co-information for this system is -0.5 bits. This example illustrates how interpreting the pointwise co-information terms allows us to select only those representing redundancy.

Table 5. Pointwise values from $I_{\text{ccs}}(S; \{1\}\{2\})$ for SUM.

(x_1, x_2, s)	$\Delta_s h(x_1)$	$\Delta_s h(x_2)$	$\Delta_s h(x_1, x_2)$	$c(x_1; x_2; s)$	$\Delta_s h^{\text{com}}(x_1, x_2)$
$(0,0,0)$	1	1	2	0	0
$(0,1,1)$	0	0	1	-1	0
$(1,0,1)$	0	0	1	-1	0
$(1,1,2)$	1	1	2	0	0

4.3. Operational Motivation for Choice of Joint Distribution

4.3.1. A Game-Theoretic Operational Definition of Unique Information

Bertschinger et al. [12] introduce an operational interpretation of unique information based on decision theory, and use that to argue the "unique and shared information should only depend on the marginal [source-target] distributions" $P(\mathbf{A_i}, S)$ (their Assumption (*) and Lemma 2). Under the assumption that those marginals alone should specify redundancy they find I_{broja} via maximisation of co-information. Here we review and extend their operational argument and arrive at a different conclusion.

Bertschinger et al. [12] operationalise unique information based on the idea that if an agent, Alice, has access to unique information that is not available to a second agent, Bob, there should be some situations in which Alice can exploit this information to gain a systematic advantage over Bob ([7], Appendix B therein). They formalise this as a decision problem, with the systematic advantage corresponding to a higher expected reward for Alice than Bob. They define a decision problem as a tuple $(p, \mathcal{A}, u)$ where $p(S)$ is the marginal distribution of the target, S, $\mathcal{A}$ is a set of possible actions the agent can take, and $u(s, a)$ is the reward function specifying the reward for each $s \in S$ $a \in \mathcal{A}$. They assert that unique information exists if and only if there exists a decision problem in which there is higher expected reward for an agent making optimal decisions based on observation of X_1, versus an agent making optimal decisions on observations of X_2. This motivates their fundamental assumption that unique information depends only on the pairwise target-predictor marginals $P(X_1, S)$, $P(X_2, S)$ ([12] Assumption *), and their assertion that $P(X_1, S) = P(X_2, S)$ implies no unique information in either predictor.

We argue that the decision problem they consider is too restrictive, and therefore the conclusions they draw about the properties of unique and redundant information are incorrect. Those properties come directly from the structure of the decision problem; the reward function u is the same for both agents, and the agents play independently from one other. The expected reward is calculated separately for each agent, ignoring by design any trial by trial covariation in their observed evidence $P(X_1, X_2)$, and resulting actions.

While it is certainly true that if their decision problem criterion is met, then there is unique information, we argue that the decision problem advantage is not a necessary condition for the existence of unique information. We prove this by presenting below a counter-example, in which we demonstrate unique information without a decision theoretic advantage. To construct this example, we extend their argument to a game-theoretic setting, where we explicitly consider two agents playing against each other. Decision theory is usually defined as the study of individual agents, while situations with multiple interacting agents are the purview of game theory. Since the unique information setup includes two agents, it seems more natural to use a game theoretic approach. Apart from switching from a decision theoretic to a game theoretic perspective, we make exactly the same argument. It is possible to operationalise unique information so that unique information exists if and only if there exists a game (with certain properties described below) where one agent obtains a higher expected reward when both agents are playing optimally under the same utility function.

We consider two agents interacting in a game, specifically a non-cooperative, simultaneous, one-shot game [40] where both agents have the same utility function. Non-cooperative means the players cannot form alliances or agreements. Simultaneous (as opposed to sequential) means the players move simultaneously; if not actually simultaneous in implementation such games can be effectively simultaneous as long as each player is not aware of the other players actions. This is a crucial requirement for a setup to operationalise unique information because if the game was sequential, it would be possible for information to "leak" from the first players evidence, via the first players action, to the second. Restricting to simultaneous games prevents this, and ensures each game provides a fair test for unique information in each players individual predictor evidence. One-shot (as opposed to repeated) means the game is played only once as a one off, or at least each play is completely independent of any other. Players have no knowledge of previous iterations, or opportunity to learn from or adapt to the actions of the other player. The fact that the utility function is the same for the actions of each player makes it a fair test for any advantage given by unique information—both players are playing by the same rules. These requirements ensure that, as for the decision theoretic argument of [12], each player must chose an action to maximise their reward based only the evidence they observe from the predictor variable. If a player is able to obtain a systematic advantage, in the form of a higher expected reward for some specific game, given the game is fair and they are acting only on the information in the predictor they observe, then this must correspond to unique information in that predictor. This is the same as the claim made in [12] that higher expected reward in a specific decision problem implies unique information in the predictor.

In fact, if in addition to the above properties the considered game is also symmetric and non-zero-sum then this is exactly equivalent to the decision theoretic formulation. Symmetric means the utility function is invariant to changes of player identity (i.e., it is the same if the players swap places). Alternatively, an asymmetric game is one in which the reward is not necessarily unchanged if the identity of the players is switched. A zero-sum game is one in which there is a fixed reward that is distributed between the players while in a non-zero-sum game the reward is not fixed. The decision problem setup is non-zero-sum, since the action of one agent does not affect the reward obtained by the other agent. Both players consider the game as a decision problem and so play as they would in the decision theoretic framework (i.e., to choose an action based only on their observed evidence in such a way as to maximise their expected reward). This is because since the game is non-cooperative, simultaneous and one-shot they have no knowledge of or exposure to the other players actions.

We argue unique information should also be operationalised in asymmetric and zero-sum games, since these also satisfy the core requirements outlined above for a fair test of unique information. In a zero-sum game, the reward of each agent now also depends on the action of the other agent, therefore unique information is not invariant to changes in $P(X_1, X_2)$, because this can change the balance of rewards on individual realisations. Note that this does not require either player is aware of the others actions (because the game is simultaneous), they still chose an action based only on their own predictor evidence, but their reward depends also on the action of the other agent (although those actions themselves are invisible). The stochastic nature of the reward from the perspective of each individual agent is not an issue since, as for the decision theoretic approach, we consider only one-shot games. Alternatively, if an asymmetry is introduced to the game, for example by allowing one agent to set the stake in a gambling task, then again $P(X_1, X_2)$ affects the unique information. We provide a specific example for this second case, and specify an actual game which meets the above requirements and provides a systematic advantage to one player, demonstrating the presence of unique information. However, this system does not admit a decision problem which provides an advantage. This counter-example therefore proves that the decision theoretic operationalisation of [12] is not a necessary condition for the existence of unique information.

Borrowing notation from [12] we consider two agents, which each observe values from X_1 and X_2 respectively, and take actions $a_1, a_2 \in \mathcal{A}$. Both are subject the same core utility function $v(s,a)$, but we break the symmetry in the game by allowing one agent to perform a second action—setting the stake on each hand (realisation). This results in utility functions $u_i(s, a_i, x_1) = c(x_1)v(s, a_i)$, where c is a stake weighting chosen by agent 1 on the basis of their evidence. This stake weighting is not related to their guess on the value s (their action a_i), but serves here as a way to break the symmetry of the game while maintaining equal utility functions for each player. That is, although the reward here is a function also of x_1, it is the same function for both players, so $a_1 = a_2 \implies u_1(s, a_1, x_1) = u_2(s, a_2, x_1) \forall s, x_1$. In general, in the game theoretic setting the utility function can depend on the entire state of the world, $u(s, a_i, x_1, x_2)$, but here we introduce only an asymmetric dependence on x_1. Both agents have the same utility function as required for a fair test of unique information, but that utility function is asymmetric—it is not invariant to switching the players. The second agent is not aware of the stake weighting applied to the game when they choose their action. The tuple $(p, \mathcal{A}, \mathbf{u})$ defines the game with $\mathbf{u}(s, a_1, a_2, x_1) = [u_1(s, a_1, x_1), u_2(s, a_2, x_1)]$. In this case the reward of agent 2 depends on x_1, introducing again a dependence on $P(X_1, X_2)$. However, because both agents have the same asymmetric utility function, this game meets the intuitive requirements for an operational test of unique information. If there is no unique information, agent 1 should not be able to profit simply by changing the stakes on different trials. If they can profit systematically by changing the stakes on trials that are favourable to them based on the evidence they observe, that is surely an operationalisation of unique information. We emphasise again that we are considering here a non-cooperative, simultaneous, one-shot, non-zero-sum, asymmetric game. So agent 2 does not have any information about the stake weight on individual games, and cannot learn anything about the stake weight from repeated plays. Therefore, there is no way for unique information in X_1 to affect the action of agent 2 via the stake weight setting. The only difference from the decision theoretic framework is that here we consider an asymmetric utility function.

To demonstrate this, and provide a concrete counter-example to the decision theoretic argument [12] we consider a system termed REDUCEDOR (Joseph Lizier, *personal communication*). Figure 4A shows the probability distribution which defines this binary system. Table 6 shows the PIDs for this system. Figure 4B shows the distribution resulting from the I_{broja} optimisation procedure. Both systems have the same target-predictor marginals $P(X_i, S)$, but have different predictor-predictor marginals $P(X_1, X_2)$. I_{broja} reports zero unique information. I_{ccs} reports zero redundancy, but unique information present in both predictors.

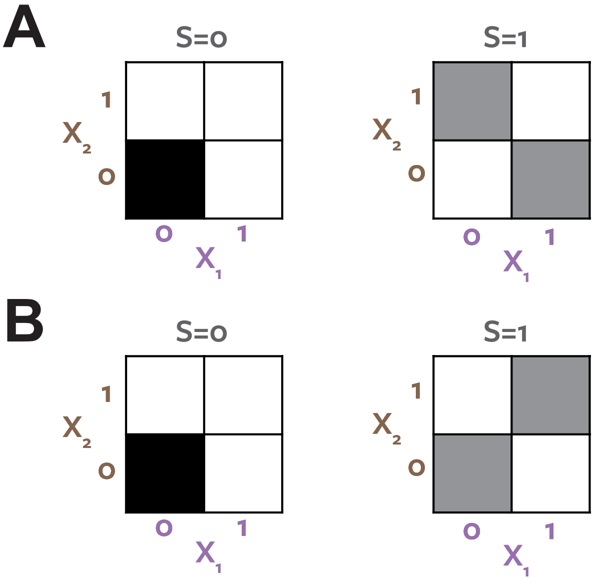

Figure 4. REDUCEDOR. (**A**) Probability distribution of REDUCEDOR system; (**B**) Distribution resulting from I_{broja} optimisation. Black tiles represent outcomes with $p = 0.5$. Grey tiles represent outcomes with $p = 0.25$. White tiles are zero-probability outcomes.

Table 6. Partial Information Decompositions (PIDs) for REDUCEDOR (Figure 4A).

Node	$I_\partial[I_{\text{min}}]$	$I_\partial[I_{\text{broja}}]$	$I_\partial[I_{\text{ccs}}]$
{1}{2}	0.31	0.31	0
{1}	0	0	0.31
{2}	0	0	0.31
{12}	0.69	0.69	0.38

In the I_{broja} optimised distribution (Figure 4B) the two predictors are directly coupled, $P(X_1 = 0, X_2 = 1) = P(X_1 = 1, X_2 = 0) = 0$. In this case there is clearly no unique information. The coupled marginals mean both agents see the same evidence on each realisation, make the same choice and therefore obtain the same reward, regardless of the stake weighting chosen by agent 1. However, in the actual system, the situation is different. Now the evidence is de-coupled, the agents never both see the evidence $x_i = 1$ on any particular realisation $P(X_1 = 1, X_2 = 1) = 0$. Assuming a utility function $v(s, a) = \delta_{sa}$ reflecting a guessing game task, the optimal strategy for both agents is to make a guess $a_i = 0$ when they observe $x_i = 0$, and guess $a_i = 1$ when they observe $x_i = 1$. If Alice (X_1) controls the stake weight she can choose $c(x_1) = 1 + x_1$ which results in a doubling of the reward when she observes $X_1 = 1$ versus when she observes $X_1 = 0$. Under the true distribution of the system for realisations where $x_1 = 1$, we know that $x_2 = 0$ and $s = 1$, so Bob will guess $a_2 = 0$ and be wrong (have zero reward). On an equal number of trials Bob will see $x_2 = 1$, guess correctly and Alice will win nothing, but those trials have half the utility of the trials that Alice wins due to the asymmetry resulting from her specifying the gambling stake. Therefore, on average, Alice will have a systematically higher reward as a result of exploiting her unique information, which is unique because on specific realisations it is available only to her. Similarly, the argument can be reversed, and if Bob gets to choose the stakes, corresponding to a utility weighting $c(x_2) = 1 + x_2$, he can exploit unique information available to him on a separate set of realisations.

Both games considered above would provide no advantage when applied to the I_{broja} distribution (Figure 4B). The information available to each agent when they observe $X_i = 1$ is not unique, because it always occurs together on the same realisations. There is no way to gain an advantage in any game since it will always be available simultaneously to the other agent. In both decompositions the information corresponding to prediction of the stimulus when $x_i = 1$ is quantified as 0.31 bits. I_{broja} quantifies this as redundancy because it ignores the structure of $P(X_1, X_2)$ and so does not consider the within trial relationships between the agents evidence. I_{broja} cannot distinguish between the two distributions illustrated in Figure 4. I_{ccs} quantifies the 0.31 bits as unique information in both predictors, because in the true system each agent sees the informative evidence on different trials, and so can exploit it to gain a higher reward in a certain game. I_{ccs} agrees with I_{broja} in the system in Figure 4B, because here the same evidence is always available to both agents, so is not unique.

We argue that this example directly illustrates the fact that unique information is not invariant to $P(X_1, X_2)$, and that the decision theoretic operational definition of [12] is too restrictive. The decision theory view says that unique information corresponds to an advantage which can be obtained only when two players go to different private rooms in a casino, play independently and then compare their winnings at the end of the session. The game theoretic view says that unique information corresponds to any obtainable advantage in a fair game (simultaneous and with equal utility functions), even when the players play each other directly, betting with a fixed pot, on the same hands at the same table. We have shown a specific example where there is an advantage in the second case, but not the first case. We suggest such an advantage cannot arise without unique information in the predictor and therefore claim this counter-example proves that the decision theoretic operationalisation is not a necessary condition for the existence of unique information. While this is a single specific system, we will see in the examples (Section 5) that the phenomenon of I_{broja} over-stating redundancy by neglecting unique information which is masked when the inputs are coupled occurs frequently. We argue this occurs because the I_{broja} optimisation maximises co-information. It therefore couples the predictors to maximise the contribution of source redundancy to the co-information, since the game theoretic operationalisation shows that redundancy is not invariant to the predictor-predictor marginal distribution.

4.3.2. Maximum Entropy Optimisation

For simplicity we consider first a two-predictor system. The game-theoretic operational definition of unique information provided in the previous section requires that the unique information (and hence redundancy) should depend only on the pairwise marginals $P(S, X_1)$, $P(S, X_2)$ and $P(X_1, X_2)$. Therefore, any measure of redundancy which is consistent with this operational definition should take a constant value over the family of distributions which satisfy those marginal constraints. This is the same argument applied in [12] but we consider here the game-theoretic extension to their decision theoretic operationalisation. Co-information itself is not constant over this family of distributions, because its value can be altered by third order interactions (i.e., those not specified by the pairwise marginals). Consider for example XOR. The co-information of this distribution is -1 bits, but the maximum entropy distribution preserving pairwise marginal constraints is the uniform distribution with a co-information of 0 bits. Therefore, if I_{ccs} were calculated using the full joint distribution it would not be consistent with the game-theoretic operational definition of unique information.

Since redundancy should be invariant given the specified marginals, our definition of I_{ccs} must be a function only of those marginals. However, we need a full joint distribution over the trivariate joint space to calculate the pointwise co-information terms. We use the maximum entropy distribution subject to the constraints specified by the game-theoretic operational definition (Equation (31)). The maximum entropy distribution is by definition the most parsimonious way to fill out a full trivariate distribution given only a set of bi-variate marginals [41]. It introduces no additional structure to the 3-way distribution over that which is specified by the constraints. Pairwise marginal constrained maximum entropy distributions have been widely used to study the effect of third and higher order interactions, because they provide a surrogate model which removes these effects [42–45]. Any distribution with lower entropy would by definition have some additional structure over that which is required to specify the unique and redundant information following the game-theoretic operationalisation.

Note that the definition of I_{broja} follows a similar argument. If redundancy was measured with co-information directly, it would not be consistent with the decision theoretic operationalisation [12]. Bertschinger et al. [12] address this by choosing the distribution which maximises co-information subject to the decision theoretic constraints. While we argue that maximizing entropy is in general a more principled approach than maximizing co-information, note that with the additional predictor marginal constraint introduced by the game-theoretic operational definition, both approaches are equivalent for two predictors (since maximizing co-information is equal to maximizing entropy given the constraints). However, once the distribution is obtained the other crucial difference is that I_{ccs} separates genuine redundant contributions at the local level, while I_{broja} computes the full co-information, which conflates redundant and synergistic effects (Table 3) [6].

We apply our game-theoretic operational definition in the same way to provide the constraints in Equation (31) for an arbitrary number of inputs. The action of each agent is determined by $P(A_i, S)$ (or equivalently $P(S|A_i)$) and the agent interaction effects (from zero-sum or asymmetric utility functions) are determined by $P(A_1, \ldots, A_n)$.

4.4. Properties

The measure I_{ccs} as defined above satisfies some of the proposed redundancy axioms (Section 3). The symmetry and self-redundancy axioms are satisfied from the properties of co-information [20]. For self-redundancy, consider that co-information for $n = 2$ is equal to mutual information at the pointwise level (Equation (29)):

$$\begin{aligned} c(s,a) &= h(s) + h(a) - h(s,a) \\ &= i(s;a) = \Delta_s h(a) \end{aligned} \tag{38}$$

So $\operatorname{sgn} c(s,a) = \operatorname{sgn} \Delta_s h(a) \;\; \forall s, a$ and $I_{\text{ccs}}(S;A) = I(S;A)$. Subset equality is also satisfied. If $\mathbf{A_{l-1}} \subseteq \mathbf{A_l}$ then we consider values $a_{l-1} \in \mathbf{A_{l-1}}$, $a_l \in \mathbf{A_l}$ with $a_l = (a_l^{l-1}, a_l^+)$ and $a_l^{l-1} \in \mathbf{A_{l-1}} \cap \mathbf{A_l} = \mathbf{A_{l-1}}$, $a_l^+ \in \mathbf{A_l} \setminus \mathbf{A_{l-1}}$. Then

$$p(a_{i_1}, \ldots, a_{i_j}, a_{l-1}, a_l^{l-1}, a_l^+) = \begin{cases} 0 & \text{if } a_{l-1} \neq a_l^{l-1} \\ p(a_{i_1}, \ldots, a_{i_j}, a_l) & \text{otherwise} \end{cases} \tag{39}$$

for any $i_1 < \cdots < i_j \in \{1, \ldots, l-2\}$. So for non-zero terms in Equation (30):

$$h(a_{i_1}, \ldots, a_{i_j}, a_{l-1}, a_l) = h(a_{i_1}, \ldots, a_{i_j}, a_l) \tag{40}$$

Therefore all terms for $k \geq 2$ in Equation (29) which include a_{l-1}, a_l cancel with a corresponding $k-1$ order term including a_l, so

$$c(a_1, \ldots, a_{l-1}, a_l) = c(a_1, \ldots, a_{l-1}) \tag{41}$$

and subset equality holds.

I_{ccs} does not satisfy the Harder et al. identity axiom [11] (Equation (19)); any distribution with negative local information terms serves as a counter example. These negative terms represent synergistic entropy which is included the standard mutual information quantity [36]. Therefore their omission in the calculation of I_{ccs} seems appealing; since they result from a synergistic interaction they should not be included in a measure quantifying redundant information. I_{ccs} does satisfy the modified independent identity axiom (Equation (22)), and so correctly quantifies redundancy in the two-bit copy problem (Section 3.2).

However, I_{ccs} does not satisfy monotonicity. To demonstrate this, consider the following example (Table 7, modified from [13], Figure 3).

Table 7. Example system with unique misinformation.

x_1	x_2	s	$p(x_1, x_2, s)$
0	0	0	0.4
0	1	0	0.1
1	1	1	0.5

For this system,

$$I(S;X_1) = I(S;X_1,X_2) = 1 \text{ bit}$$
$$I(S;X_2) = 0.61 \text{ bits}$$

Because of the self redundancy property, these values specify $I_\cap$ for the upper 3 values of the redundancy lattice (Figure 2A). The value of the bottom node is given by

$$I_\partial = I_\cap = I_{\text{ccs}}(S;\{1\}\{2\}) = 0.77 \text{ bits}$$

This value arises from two positive pointwise terms:

$$x_1 = x_2 = s = 0 \text{ (contributes 0.4 bits)}$$
$$x_1 = x_2 = s = 1 \text{ (contributes 0.37 bits)}$$

So $I_{\text{ccs}}(S;\{1\}\{2\}) > I_{\text{ccs}}(S;\{2\})$ which violates monotonicity on the lattice. How is it possible for two variables to share more information than one of them carries alone?

Consider the pointwise mutual information values for $I_{\text{ccs}}(S;\{2\}) = I(S;X_2)$. There are the same two positive information terms that contribute to the redundancy (since both are common with X_1). However, there is also a third misinformation term of -0.16 bits when $s = 0, x_2 = 1$. In our view, this demonstrates that the monotonicity axiom is incorrect for a measure of redundant information content. As this example shows a node can have *unique misinformation*.

For this example I_{ccs} yields the PID:

$$I_\partial(\{1\}\{2\}) = 0.77$$
$$I_\partial(\{1\}) = 0.23$$
$$I_\partial(\{2\}) = -0.16$$
$$I_\partial(\{12\}) = 0.16$$

While monotonicity has been considered a crucial axiom with the PID framework, we argue that subset equality, usually considered as part of the axiom of monotonicity, is the essential property that permits the use of the redundancy lattice. We have seen this lack of monotonicity means the PID obtained with I_{ccs} is not non-negative. We agree that while “negative … atoms can *subjectively* be seen as flaw" [37], we argue here that in fact they are a necessary consequence of a redundancy measure that genuinely quantifies overlapping information content. Please note that in an earlier version of this manuscript we proposed thresholding with 0 to remove negative values. We no longer do so.

Mutual information is the expectation of a local quantity that can take both positive (local information) and negative (local misinformation) values, corresponding to redundant and synergistic entropy respectively [36]. Jensen's equality ensures that the final expectation value of mutual information is positive; or equivalently that redundant entropy is greater than synergistic entropy in any bivariate system. We argue that when breaking down the classical Shannon information into a partial information decomposition, there is no reason that those partial information values must be non-negative, since there is no way to apply Jensen's inequality to these partial values. We have illustrated this with a simple example where a negative unique information value is obtained, and inspection of the pointwise terms shows that this is indeed due to negative pointwise terms in the mutual information calculation for one predictor that are not present in the mutual information calculation for the other predictor: unique misinformation. Applying the redundancy lattice and the partial information decomposition directly to entropy can provide some further insights into the prevalence and effects of misinformation or synergistic entropy [36].

We conjecture that I_{ccs} is continuous in the underlying probability distribution [46] from the continuity of the logarithm and co-information, but not differentiable due to the thresholding with 0. Continuity requires that, at the local level,

$$c(s, a_1, a_2) < \min\left[i(s;a_1), i(s;a_2)\right] \tag{42}$$

when $\operatorname{sgn} i(s;a_1) = \operatorname{sgn} i(s;a_2) = \operatorname{sgn} i(s;a_1,a_2) = \operatorname{sgn} c(s,a_1,a_2)$. While this relationship holds for the full integrated quantities [20], it does not hold at the local level for all joint distributions. However, we conjecture that it holds when using the pairwise maximum entropy solution $\hat{P}$, with no higher order interactions. This is equivalent to saying that the overlap of the two local informations should not be larger than the smallest—an intuitive requirement for a set theoretic overlap. However, at this stage the claim of continuity remains a conjecture. In the Matlab implementation we explicitly test for violations of the condition in Equation (42), which do not occur in any of the examples we consider here. This shows that all the examples we consider here are at least locally continuous in the neighbourhood of the specific joint probability distribution considered.

In the next sections, we demonstrate with a range of example systems how the results obtained with this approach match intuitive expectations for a partial information decomposition.

4.5. Implementation

Matlab code is provided to accompany this article, which features simple functions for calculating the partial information decomposition for two and three variables [47].This includes implementation of I_{min} and the PID calculation of [6], as well as I_{ccs} and I_{broja}. Scripts are provided reproducing all the examples considered here. Implementations of I_{ccs} and I_{mmi} [26] for Gaussian systems are also included. To calculate I_{broja} and compute the maximum entropy distributions under marginal constraints we use the `dit` package [48–50]

5. Two Variable Examples

5.1. Examples from Williams and Beer (2010) [6]

We begin with the original examples of ([6], Figure 4), reproduced here in Figure 5.

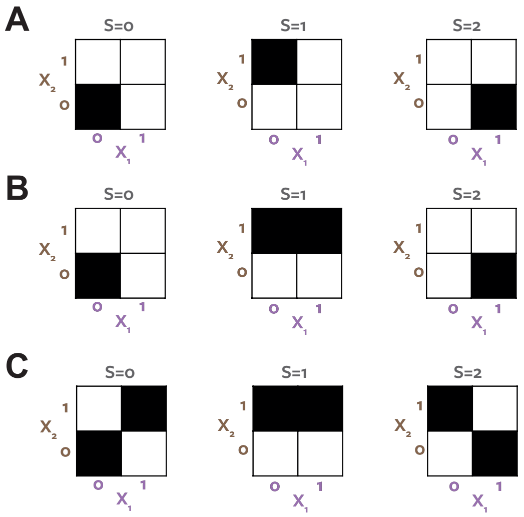

Figure 5. Probability distributions for three example systems (**A**–**C**). Black tiles represent equiprobable outcomes. White tiles are zero-probability outcomes. (**A**,**B**) modified from [6].

Table 8 shows the PIDs for the system shown in Figure 5A, obtained with I_{min}, I_{broja} and I_{ccs}. Note that this is equivalent to the system SUBTLE in ([13], Figure 4). I_{ccs} and I_{min} agree qualitatively here; both show both synergistic and redundant information. I_{broja} shows zero synergy. The pointwise computation of I_{ccs} includes two non-zero terms; when

$$x_1 = 0, x_2 = 1, s = 1 \text{ and when}$$
$$x_1 = 1, x_2 = 0, s = 2$$

For both of these local values, x_1 and x_2 are contributing the same reduction in surprisal of s (0.195 bits each for 0.39 bits overall redundancy). There are no other redundant local changes in surprisal (positive or negative). In this case, both the I_{broja} optimised distribution and the pairwise marginal maximum entropy distribution are equal to the original distribution. So here I_{broja} is measuring redundancy directly with co-information, whereas I_{ccs} breaks down the co-information to include only the two terms which directly represent redundancy. In the full co-information calculation of I_{broja} there is one additional contribution of -0.138 bits, which comes from the $x_1 = x_2 = s = 0$ event. In this case the local changes in surprisal of s from x_1 and x_2 are both positive (0.585), but the local co-information is negative (-0.415). This corresponds to the second row of Table 3—it is synergistic local information. Therefore this example clearly shows how the I_{broja} measure of redundancy erroneously includes synergistic effects.

Table 8. PIDs for example Figure 5A.

Node	$I_\partial[I_{\text{min}}]$	$I_\partial[I_{\text{broja}}]$	$I_\partial[I_{\text{ccs}}]$
{1}{2}	0.5850	0.2516	0.3900
{1}	0.3333	0.6667	0.5283
{2}	0.3333	0.6667	0.5283
{12}	0.3333	0	0.1383

Table 9 shows the PIDs for the system shown in Figure 5B. Here I_{broja} and I_{ccs} agree, but diverge qualitatively from I_{min}. I_{min} shows both synergy and redundancy, with no unique information carried by X_1 alone. I_{ccs} shows no synergy and redundancy, only unique information carried independently by X_1 and X_2. Reference [6] argue that "X_1 and X_2 provide 0.5 bits of redundant information corresponding to the fact that knowledge of either X_1 or X_2 reduces uncertainty about the outcomes

$S = 0, S = 2$". However, while both variables reduce uncertainty about S, they do so in different ways—X_1 discriminates the possibilities $S = 0, 1$ vs. $S = 1, 2$ while X_2 allows discrimination between $S = 1$ vs. $S = 0, 2$. These discriminations represent different non-overlapping information content, and therefore should be allocated as unique information to each variable as in the I_{ccs} and I_{broja} PIDs. While the full outcome can only be determined with knowledge of both variables, there is no synergistic information because the discriminations described above are independent.

Table 9. PIDs for example Figure 5B.

Node	$I_\partial[I_{\text{min}}]$	$I_\partial[I_{\text{broja}}]$	$I_\partial[I_{\text{ccs}}]$
$\{1\}\{2\}$	0.5	0	0
$\{1\}$	0	0.5	0.5
$\{2\}$	0.5	1	1
$\{12\}$	0.5	0	0

To induce genuine synergy it is necessary to make the X_1 discrimination between $S = 0, 1$ and $S = 1, 2$ ambiguous without knowledge of X_2. Table 10 shows the PID for the system shown in Figure 5C, which includes such an ambiguity. Now there is no information in X_1 alone, but it contributes synergistic information when X_2 is known. Here, I_{min} correctly measures 0 bits redundancy, and all three PIDs agree (the other three terms have only one source, and therefore are the same for all measures from self-redundancy).

Table 10. PIDs for example Figure 5C.

Node	$I_\partial[I_{\text{min}}]$	$I_\partial[I_{\text{broja}}]$	$I_\partial[I_{\text{ccs}}]$
$\{1\}\{2\}$	0	0	0
$\{1\}$	0	0	0
$\{2\}$	0.25	0.25	0.25
$\{12\}$	0.67	0.67	0.67

5.2. Binary Logical Operators

The binary logical operators OR, XOR and AND are often used as example systems [10–12]. For XOR, the I_{ccs} PID agrees with both I_{min} and I_{broja} and quantifies the 1 bit of information as fully synergistic.

5.2.1. AND/OR

Figure 6 illustrates the probability distributions for AND and OR. This makes clear the equivalence between them; because of symmetry any PID should give the same result on both systems. Table 11 shows the PIDs. In this system I_{min} and I_{broja} agree, both showing no unique information. I_{ccs} shows less redundancy, and unique information in both predictors. The redundancy value with I_{ccs} falls within the bounds proposed in ([10], Figure 6.11).

To see where this unique information arises with I_{ccs} we can consider directly the individual pointwise contributions for the AND example (Table 12). $I_{\text{ccs}}(\{1\}\{2\})$ has a single pointwise contribution from the event $(0, 0, 0)$, only when both inputs are 0 is there redundant local information about the outcome. For the event $(0, 1, 0)$ (and symmetrically for $1, 0, 0$) x_1 conveys local information about s, while x_2 conveys local misinformation, therefore there is no redundancy, but a unique contribution for both x_1 and x_2. We can see in the $(1, 1, 1)$ event the change in surprisal of s from the two predictors is independent, so again contributes unique rather than redundant information. So the unique information in each predictor is a combination of unique information and misinformation terms.

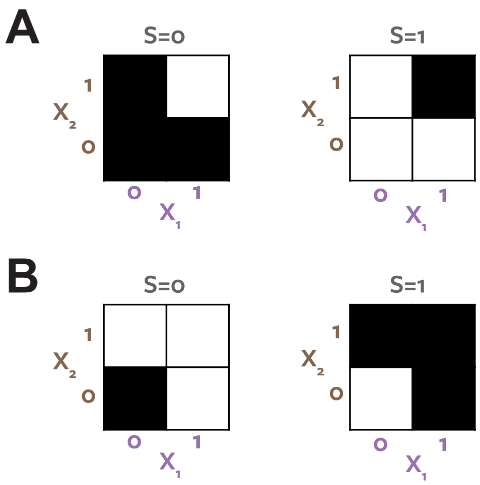

Figure 6. Binary logical operators. Probability distributions for (**A**) AND; (**B**): OR. Black tiles represent equiprobable outcomes. White tiles are zero-probability outcomes.

Table 11. PIDs for AND/OR.

Node	$I_\partial[I_{\text{min}}]$	$I_\partial[I_{\text{broja}}]$	$I_\partial[I_{\text{ccs}}]$
{1}{2}	0.31	0.31	0.10
{1}	0	0	0.21
{2}	0	0	0.21
{12}	0.5	0.5	0.29

Table 12. Pointwise values from $I_{\text{ccs}}(S;\{1\}\{2\})$ for AND.

(x_1, x_2, s)	$\Delta_s h(x_1)$	$\Delta_s h(x_2)$	$\Delta_s h(x_1, x_2)$	$c(x_1; x_2; s)$	$\Delta_s h^{\text{com}}(x_1, x_2)$
$(0, 0, 0)$	0.415	0.415	0.415	0.415	0.415
$(0, 1, 0)$	0.415	−0.585	0.415	−0.585	0
$(1, 0, 0)$	−0.585	0.415	0.415	−0.585	0
$(1, 1, 1)$	1	1	2	0	0

For I_{broja} the specific joint distribution that maximises the co-information in the AND example while preserving $P(X_i, S)$ ([12], Example 30, $\alpha = 1/4$) has an entropy of 1.5 bits. $\hat{P}(X_1, X_2, S)$ used in the calculation of I_{ccs} is equal to the original distribution and has an entropy of 2 bits. Therefore, the distribution used in I_{broja} has some additional structure above that specified by the individual joint target marginals and which is chosen to maximise the co-information (negative interaction information). As discussed above, interaction information can conflate redundant information with synergistic misinformation, as well as having other ambiguous terms when the signs of the individual changes of surprisal are not equal. As shown in Table 12, the AND system includes such ambiguous terms (rows 2 and 3, which contribute synergy to the interaction information). Any system of the form considered in ([12], Example 30) will have similar contributing terms. This illustrates the problem with using co-information directly as a redundancy measure, regardless of how the underlying distribution is obtained. The distribution selected to maximise co-information will be affected by these ambiguous and synergistic terms. In fact, it is interesting to note that for the I_{broja} distribution ($\alpha = 1/4$), $p(0, 1, 0) = p(1, 0, 0) = 0$ and the two ambiguous synergistic terms are removed from the interaction information. This indicates how the optimisation of the co-information might be driven by terms that cannot be interpreted as genuine redundancy. Further, the distribution used in I_{broja} has perfectly coupled marginals. This increases the source redundancy measured by the co-information. Under this distribution, the $(1, 1, 1)$ term now contributes 1 bit locally to the co-information. This is redundant because $x_1 = 1$ and $x_2 = 1$ always occur together. In the original distribution the $(1, 1, 1)$ term is independent because the predictors are independent.

We argue there is no fundamental conceptual problem with the presence of unique information in the AND example. Both variables share some information, have some synergistic information, but also have some unique information corresponding to the fact that knowledge of either variable taking the value 1 reduces the uncertainty of $s = 1$ independently (i.e., on different trials). If the joint target marginal distributions are equal, then by symmetry $I_\partial(\{1\}) = I_\partial(\{2\})$, but it is not necessary that $I_\partial(\{1\}) = I_\partial(\{2\}) = 0$ ([12], Corollary 8).

5.2.2. SUM

While not strictly a binary logic gate, we also consider the summation of two binary inputs. The AND gate can be thought of as a thresholded version of summation. Summation of two binary inputs is also equivalent to the system XORAND [10–12]. Table 13 shows the PIDs.

Table 13. PIDs for SUM.

Node	$I_\partial[I_{\text{min}}]$	$I_\partial[I_{\text{broja}}]$	$I_\partial[I_{\text{ccs}}]$
$\{1\}\{2\}$	0.5	0.5	0
$\{1\}$	0	0	0.5
$\{2\}$	0	0	0.5
$\{12\}$	1	1	0.5

As with AND, I_{min} and I_{broja} agree, and both allocate 0 bits of unique information. Both of these methods always allocate zero unique information when the target-predictor marginals are equal. I_{ccs} differs in that it allocates 0 redundancy. This arises for a similar reason to the differences discussed earlier for REDUCEDOR (Section 4.3). The optimised distribution used in I_{broja} has directly coupled predictors:

$$\begin{aligned} P_{\text{broja}}(X_1 = 0, X_2 = 0) = P_{\text{broja}}(X_1 = 1, X_2 = 1) = 0.5 \\ P_{\text{broja}}(X_1 = 0, X_2 = 1) = P_{\text{broja}}(X_1 = 1, X_2 = 0) = 0 \end{aligned} \tag{43}$$

While the actual system has independent uniform marginal predictors ($P(i, j) = 0.25$). In the I_{broja} calculation of co-information the local events $(0, 0, 0)$ and $(1, 1, 2)$ both contribute redundant information, because X_1 and X_2 are coupled. However, the local co-information terms for the true distribution show that the contributions of $x_1 = 0$ and $x_2 = 0$ are independent when $s = 0$ (see Table 5). Therefore, with the true distribution these contributions are actually unique information. These differences arise because of the erroneous assumption within I_{broja} that the unique and redundant information should be invariant to the predictor-predictor marginal distribution (Section 4.3). Since they are not, the I_{broja} optimisation maximises redundancy by coupling the predictors.

The resulting I_{ccs} PID seems quite intuitive. Both X_1 and X_2 each tell whether the output sum is in $(0, 1)$ or $(1, 2)$, and they do this independently, since they are distributed independently (corresponding to 0.5 bits of unique information each). However, the final full discrimination of the output can only be obtained when both inputs are observed together, providing 0.5 bits of synergy. In contrast, I_{broja} measures 0.5 bits of redundancy. It is hard to see how summation of two independent variables should be redundant as it is not apparent how two independent summands can convey overlapping information about their sum. For AND, there is redundancy between two independent inputs. I_{ccs} shows that this arises from the fact that if $x_1 = 0$ then $y = 0$ and similarly if $x_2 = 0$ then $y = 0$. So when both x_1 and x_2 are zero they are both providing the same information content—that $y = 0$, so there is redundancy. In contrast, in SUM, $x_1 = 0$ tells that $y = 0$ or $y = 1$, but which of the two particular outputs is determined independently by the values of x_2. So the information each input conveys is independent (unique) and not redundant.

5.3. Griffith and Koch (2014) Examples

Griffith and Koch [10] present two other interesting examples: RDNXOR (their Figure 6.9) and RDNUNQXOR (their Figure 6.12).

RDNXOR consists of two two-bit (4 value) inputs X_1 and X_2 and a two-bit (4 value) output S. The first component of X_1 and X_2 redundantly specifies the first component of S. The second component of S is the XOR of the second components of X_1 and X_2. This system therefore contains 1 bit of redundant information and 1 bit of synergistic information; further every value $s \in S$ has both a redundant and synergistic contribution. I_{ccs} correctly quantifies the redundancy and synergy with the PID $(1, 0, 0, 1)$ (as do both I_{min} and I_{broja}).

RDNUNQXOR consists of two three-bit (8 value) inputs X_1 and X_2 and a four-bit (16 value) output S (Figure 3). The first component of S is specified redundantly by the first components of X_1 and X_2. The second component of S is specified uniquely by the second component of X_1 and the third component of S is specified uniquely by the second component of X_2. The fourth component of S is the XOR of the third components of X_1 and X_2. Again I_{ccs} correctly quantifies the properties of the system with the PID $(1, 1, 1, 1)$, identifying the separate redundant, unique and synergistic contributions (as does I_{broja} but not I_{min}).

Note that the PID with I_{ccs} also gives the expected results for examples RND and UNQ from [10] (see example scripts in the accompanying code [47]).

5.4. Dependence on Predictor-Predictor Correlation

To directly illustrate the fundamental conceptual difference between I_{ccs} and I_{broja} we construct a family of distributions with the same target-predictor marginals and investigate the resulting decomposition as we change the predictor-predictor correlation [51].

We restrict our attention to binary variables with uniformly distributed univariate marginal distributions. We consider pairwise marginals with a symmetric dependence of the form

$$\begin{aligned} p_c(0,0) = p_c(1,1) &= {}^{(1+c)}/_4 \\ p_c(0,1) = p_c(1,0) &= {}^{(1-c)}/_4 \end{aligned} \tag{44}$$

where the parameter c specified the correlation between the two variables. We fix $c = 0.1$ for the two target-predictor marginals:

$$\begin{aligned} P(X_1, S) &= P_{0.1}(X_1, S) \\ P(X_2, S) &= P_{0.1}(X_2, S) \end{aligned} \tag{45}$$

Then with $P(X_1, X_2) = P_c(X_1, X_2)$ we can construct a trivariate joint distribution $P_c(S, X_1, X_2)$ which is consistent with these three pairwise marginals as follows [51]. This is a valid distribution for $-0.8 \leq c \leq 0.1$.

$$\begin{aligned} p_c(0,0,0) &= {}^c/_4 + {}^1/_4 \\ p_c(0,0,1) &= {}^1/_{40} - {}^c/_4 \\ p_c(0,1,0) &= {}^1/_{40} - {}^c/_4 \\ p_c(0,1,1) &= {}^c/_4 + {}^1/_5 \\ p_c(1,0,0) &= 0 \\ p_c(1,0,1) &= {}^9/_{40} \\ p_c(1,1,0) &= {}^9/_{40} \\ p_c(1,1,1) &= {}^1/_{20} \end{aligned} \tag{46}$$

Figure 7 shows I_{broja} and I_{ccs} PIDs for this system. By design the values of unique and redundant information obtained with I_{broja} do not change as a function of predictor-predictor correlation when the target-predictor marginals are fixed. With I_{ccs} the quantities change in an intuitive manner. When the predictors are positively correlated, they are redundant, when they are negatively correlated they convey unique information. When they are independent, there is an equal mix of unique and mechanistic redundancy in this system. This emphasises the different perspective also revealed in the REDUCEDOR example (Section 4.3) and the AND example (Section 5.2.1). I_{broja} reports the co-information for a distribution where the predictors are perfectly coupled. For all the values of c reported in Figure 7A, the I_{broja} optimised distribution has coupled predictor-predictor marginals:

$$\begin{aligned} P(X_1 = 0, X_2 = 1) = P(X_1 = 1, X_2 = 0) = 0 \\ P(X_1 = 0, X_2 = 0) = P(X_1 = 1, X_2 = 1) = 0.5 \end{aligned} \tag{47}$$

Therefore, I_{broja} is again insensitive to the sort of unique information that can be operationalised in a game-theoretic setting by exploiting the trial-by-trial relationships between predictors (Section 4.3).

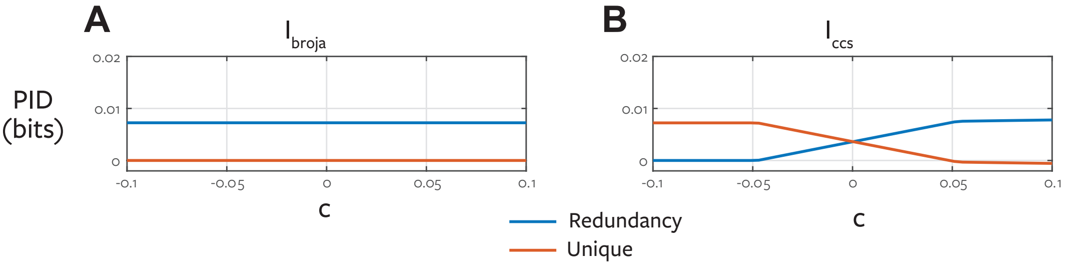

Figure 7. PIDs for binary systems with fixed target-predictor marginals as a function of predictor-predictor correlation. I_{broja} (**A**) and I_{ccs} (**B**) PIDs are shown for the system defined in Equation (46) as a function of the predictor-predictor correlation c.

6. Three Variable Examples

We now consider the PID of the information conveyed about S by three variables X_1, X_2, X_3. For three variables we do not compare to I_{broja}, since it is defined only for two input sources.

6.1. A Problem With the Three Variable Lattice?

Bertschinger et al. [14] identify a problem with the PID summation over the three-variable lattice (Figure 2B). They provide an example we term XORCOPY (described in Section 6.2.2) which demonstrates that any redundancy measure satisfying their redundancy axioms (particularly the Harder et al. identity axiom) cannot have only non-negative I_∂ terms on the lattice. We provide here an alternative example of the same problem, and one that does not depend on the particular redundancy measure used. We argue it applies for any redundancy measure that attempts to measure overlapping information content.

We consider X_1, X_2, X_3 independent binary input variables. Y is a two-bit (4 value) output with the first component given by $X_1 \oplus X_2$ and the second by $X_2 \oplus X_3$. We refer to this example as DBLXOR. In this case the top four nodes have non-zero (redundant) information:

$$\begin{aligned} I_\cap(\{123\}) = I(\{123\}) = 2 \text{ bits} \\ I_\cap(\{12\}) = I_\cap(\{13\}) = I_\cap(\{23\}) = 1 \text{ bit} \end{aligned}$$

We argue that all lower nodes on the lattice should have zero redundant (and partial) information. First, by design and from the properties of XOR no single variable conveys any information or can have any redundancy with any other source. Second, considering synergistic pairs, Figure 8A graphically illustrates the source-output joint distributions for the two-variable sources. Each value of the pairwise response (x-axes in Figure 8A) performs a different discrimination between the values of Y for each pair. Therefore, there is no way there can be redundant information between any of these synergistic pairs. Redundant information means the same information content. Since there are no discriminations (column patterns in the figure) that are common to more than one pair of sources, there can be no redundant information between them. Therefore, the information conveyed by the three two-variable sources is also independent and all lower nodes on the lattice are zero.

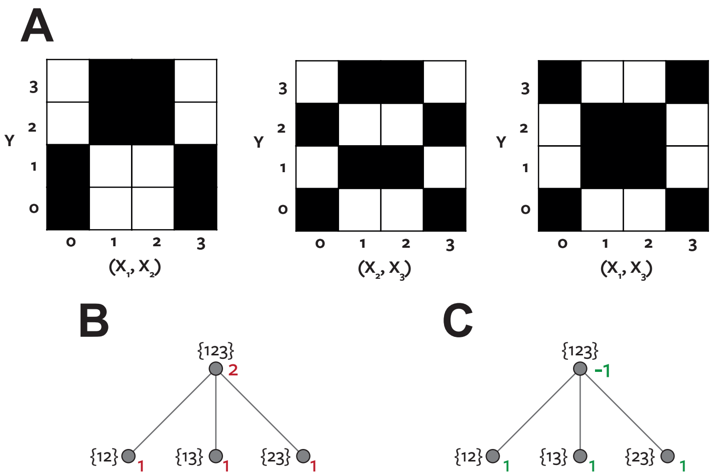

Figure 8. The DblXor example. (**A**) Pairwise variable joint distributions. Black tiles represent equiprobable outcomes. White tiles are zero-probability outcomes; (**B**) Non-zero nodes of the three variable redundancy lattice. Mutual information values for each node are shown in red; (**C**) PID. I_∂ values for each node are shown in green.

In this example, $I_\cap(\{123\}) = 2$ but there are three child nodes of $\{123\}$ each with $I_\partial = 1$ (Figure 8B). This leads to $I_\partial(\{123\}) = -1$. How can there be 3 bits of unique information in the lattice when there are only 2 bits of information in the system? In this case, we cannot appeal to the non-monotonicity of I_{ccs} since these values are monotonic on the lattice. There are also no negative pointwise terms in the calculation of $I(\{123\})$ so there is no synergistic misinformation that could explain a negative value.

In a previous version of this manuscript we argued that this problem arises because the three nodes in the penultimate level of the lattice are not disjoint, therefore not independent, and therefore mutual information is not additive over those nodes. We proposed a normalisation procedure to address such situations. However, we now propose instead to accept the negative values. As noted earlier (Section 4.4), negative values may subjectively be seen as a flaw [37], but given that mutual information itself is a summation of positive and negative terms, there is no a priori reason why a full decomposition must, or indeed can, be completely non-negative. In fact, in entropy terms, negative values are an essential consequence of the existence of mechanistic redundancy [36]. While in an information decomposition they can also arise from unique or synergistic misinformation, we propose that mechanistic redundancy is another explanation. In this particular example of DblXor, the negative $\{123\}$ term reflects a mechanistic redundancy between the three pairwise synergistic partial information terms that cannot be accounted for elsewhere on the lattice.

6.2. Other Three Variable Example Systems

6.2.1. Giant Bit and Parity

The most direct example of three-way information redundancy is the "giant bit" distribution [52]. This is the natural extension of example RDN (Section 4.2) with a single bit in common to all four variables, defined as:

$$P(0,0,0,0) = P(1,1,1,1) = 0.5 \tag{48}$$

Applying I_{ccs} results in a PID with $I_\partial(S;\{1\}\{2\}\{3\}) = 1$ bit, and all other terms zero.

A similarly classic example of synergy is the even parity distribution, a distribution in which an equal probability is assigned to all configurations with an even number of ones. The XOR distribution is the even parity distribution in the three variable (two predictor) case. Applying I_{ccs} results in a PID with $I_\partial(S;\{123\}) = 1$ bit, and all other terms zero.

Thus, the PID based on I_{ccs} correctly reflects the structure of these simple examples.

6.2.2. XORCOPY

This example was developed to illustrate the problem with the three variable lattice described above [14,53]. The system comprises three binary input variables X_1, X_2, X_3, with X_1, X_2 uniform independent and $X_3 = X_1 \oplus X_2$. The output Y is a three bit (8 value) system formed by copying the inputs $Y = (X_1, X_2, X_3)$. The PID with I_{min} gives:

$$I_\partial(\{1\}\{2\}\{3\}) = I_\partial(\{12\}\{13\}\{23\}) = 1 \text{ bit}$$

But since X_1 and X_2 are copied independently to the output it is hard to see how they can share information. Using common change in surprisal we obtain:

$$\begin{aligned} I_{\text{ccs}}(\{1\}\{23\}) = I_{\text{ccs}}(\{2\}\{13\}) = I_{\text{ccs}}(\{3\}\{12\}) &= 1 \text{ bit} \\ I_{\text{ccs}}(\{12\}\{13\}\{23\}) &= 2 \text{ bits} \end{aligned}$$

The $I_{\text{ccs}}(\{i\}\{jk\})$ values correctly match the intuitive redundancy given the structure of the system, but result in a negative value similar to DBLXOR considered above. There are 3 bits of unique I_∂ among the nodes of the third level, but only 2 bits of information in the system. This results in the PID:

$$\begin{aligned} I_\partial(\{1\}\{23\}) = I_\partial(\{2\}\{13\}) = I_\partial(\{3\}\{12\}) &= 1 \text{ bit} \\ I_\partial(\{12\}\{13\}\{23\}) &= -1 \text{ bit} \end{aligned}$$

As for DBLXOR we believe this provides a meaningful decomposition of the total mutual information, with the negative value here representing the presence of mechanistic redundancy between the nodes at the third level of the lattice. This mechanistic redundancy between synergistic pairs seems to be a signature property of an XOR mechanism.

6.2.3. Other Examples

Griffith and Koch [10] provide a number of other interesting three variable examples based on XOR operations, such as XORDUPLICATE (their Figure 6.6), XORLOSES (their Figure 6.7), XORMULTICOAL (their Figure 6.14). For all of these examples I_{ccs} provides a PID which matches what they suggest from the intuitive properties of the system (see `examples_3d.m` in accompanying code [47]). I_{ccs} also gives the correct PID for PARITYRDNRDN (which appeared in an earlier version of their manuscript).

We propose an additional example, XorUnq, which consists of three independent input bits. The output consists of 2 bits (4 values), the first of which is given by $X_1 \oplus X_2$, and the second of which is a copy of X_3. In this case we obtain the correct PID:

$$I_\partial(\{3\}) = I_\partial(\{12\}) = 1 \text{ bit}$$

Another interesting example from [10] is AndDuplicate (their Figure 6.13). In this example Y is a binary variable resulting from the binary And of X_1 and X_2. X_3 is a duplicate of X_1. The PID we obtain for this system is shown in Figure 9.

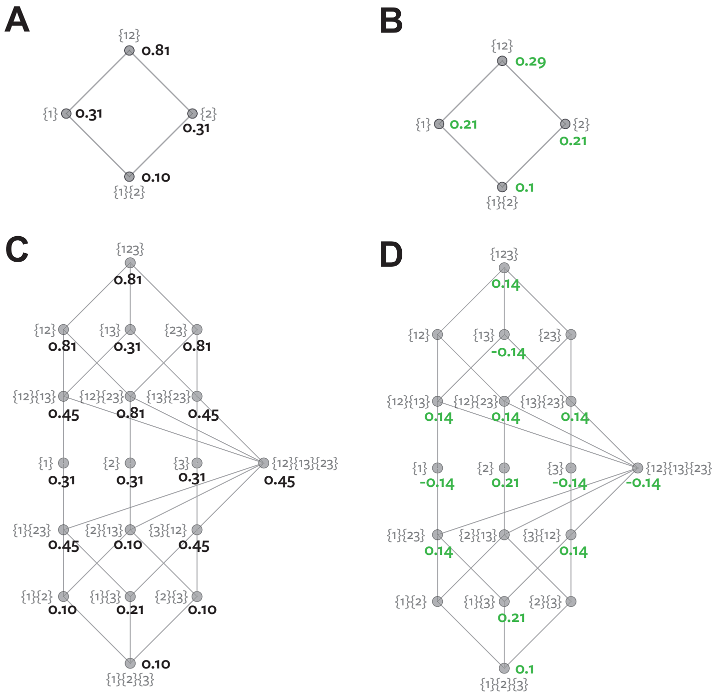

Figure 9. The AndDuplicate example. (**A**) I_{ccs} values for And; (**B**) Partial information values from the I_{ccs} PID for And; (**C**) I_{ccs} values for AndDuplicate; (**D**) Partial information values from the I_{ccs} PID for AndDuplicate.

We can see that as suggested by [10],

$$\begin{aligned} I_\partial^{\textsc{AndDup}}(S;\{2\}) &= I_\partial^{\textsc{And}}(S;\{2\}) \\ I_\partial^{\textsc{AndDup}}(S;\{1\}\{3\}) &= I_\partial^{\textsc{And}}(S;\{1\}) \\ I_\partial^{\textsc{AndDup}}(S;\{1\}\{2\}\{3\}) &= I_\partial^{\textsc{And}}(S;\{1\}\{2\}) \end{aligned} \tag{49}$$

The synergy relationship they propose, $I_\partial^{\textsc{AndDup}}(S;\{12\}\{23\}) = I_\partial^{\textsc{And}}(S;\{12\})$ is not met, although the fundamental general consistency requirement relating 2 and 3 variable lattices is [36,54]:

$$\begin{aligned} I_\partial^{\textsc{And}}(S;\{12\}) = {} & I_\partial^{\textsc{AndDup}}(S;\{12\}) \\ & + I_\partial^{\textsc{AndDup}}(S;\{12\}\{13\}) + I_\partial^{\textsc{AndDup}}(S;\{12\}\{23\}) \\ & + I_\partial^{\textsc{AndDup}}(S;\{12\}\{13\}\{23\}) \\ & + I_\partial^{\textsc{AndDup}}(S;\{3\}\{12\}) \end{aligned} \tag{50}$$

Note that the preponderance of positive and negative terms with amplitude 0.14 bits is at first glance counter-intuitive, particularly the fact that $I_\partial^{\textsc{AndDup}}(S;\{1\}) = I_\partial^{\textsc{AndDup}}(S;\{3\}) = -0.146$ when X_3 is a copy of X_1. However, the 0.14 bits comes from a local misinformation term in the univariate predictor-target mutual information calculation for AND, which is not present in the joint mutual information calculation. This reflects the fact that, in entropy terms, $I(S;X_1)$ is not a proper subset of $I(S;X_1,X_2)$ [36]. A partial entropy decomposition of AND shows that $H_\partial(\{1\}\{23\}) = H_\partial(\{2\}\{13\}) = 0.14$. These are entropy terms that have an ambiguous interpretation and appear both in unique and synergistic partial information terms. It is likely that a higher-order entropy decomposition could shed more light on the structure of the ANDDUPLICATE PID.

7. Continuous Gaussian Variables

I_{ccs} can be applied directly to continuous variables. $\Delta_s h^{\mathrm{com}}$ can be used locally in the same way, with numerical integration applied to obtain the expectation. Functions implementing this for Gaussian variables via Monte Carlo integration are included in the accompanying code [47]. Following Barrett [26] we consider the information conveyed by two Gaussian variables X_1, X_2 about a third Gaussian variable, S. We focus here on univariate Gaussians, but the accompanying implementation also supports multivariate normal distributions. Reference [26] show that for such Gaussian systems, all previous redundancy measures agree, and are equal to the minimum mutual information carried by the individual variables:

$$I_\cap(\{1\}\{2\}) = \min_{i=1,2} I(S;X_i) = I_{\mathrm{mmi}}(\{1\}\{2\}) \tag{51}$$

Without loss of generality, we consider all variables to have unit variance, and the system is then completely specified by three parameters:

$$\begin{aligned} a &= \mathrm{Corr}(X_1, S) \\ c &= \mathrm{Corr}(X_2, S) \\ b &= \mathrm{Corr}(X_1, X_2) \end{aligned}$$

Figure 10 shows the results for two families of Gaussian systems as a function of the correlation, b, between X_1 and X_2 ([26], Figure 3).

This illustrates again a key conceptual difference between I_{ccs} and existing measures. I_{ccs} is not invariant to the predictor-predictor marginal distributions (Section 5.4). When the two predictors have equal positive correlation with the target (Figure 10A,B), I_{mmi} reports zero unique information, and a constant level of redundancy regardless of the predictor-predictor correlation b. I_{ccs} transitions from having the univariate predictor information purely unique when the predictors are negatively correlated, to purely redundant when the predictors are strongly positively correlated. When the two predictors have unequal positive correlations with the target (Figure 10C,D), the same behaviour is seen. When the predictors are negatively correlated the univariate information is unique, as they become correlated both unique informations decrease as the redundancy between the predictors increases.

Having an implementation for continuous Gaussian variables is of practical importance, because for multivariate discrete systems sampling high dimensional spaces with experimental data becomes increasingly challenging. We recently developed a lower-bound approximate estimator of mutual information for continuous signals based on a Gaussian copula [3]. The Gaussian I_{ccs} measure therefore allows this approach to be used to obtain PIDs from experimental data.

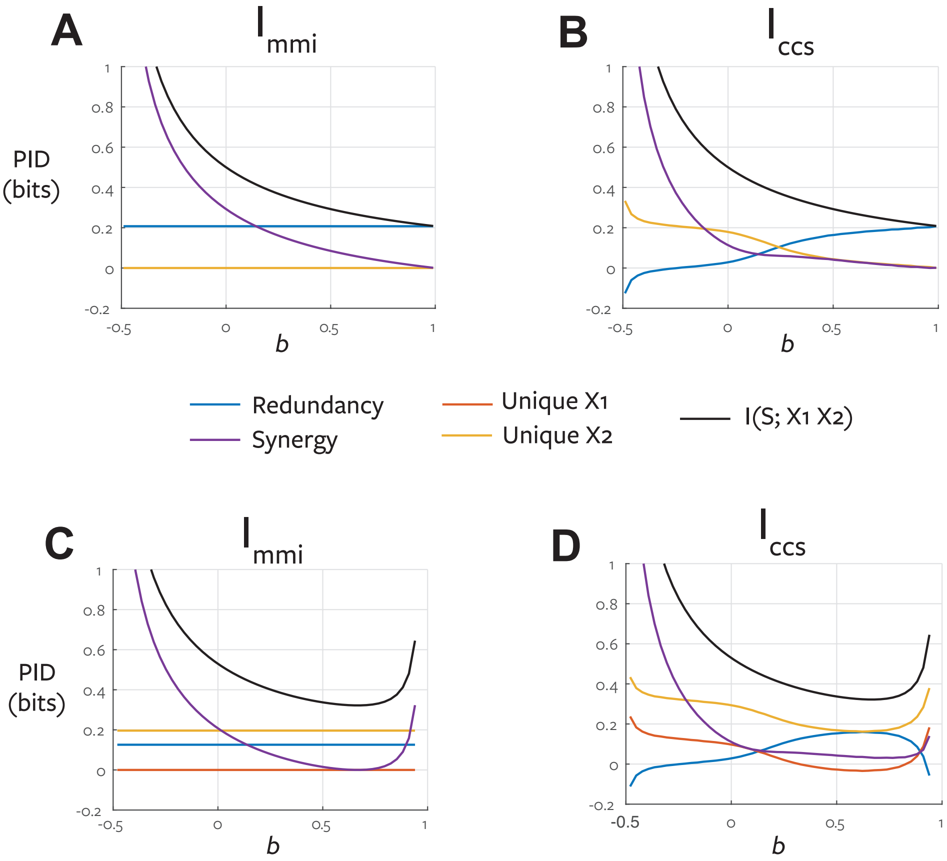

Figure 10. PIDs for Gaussian systems. (**A**) PID with I_{mmi} for $a = c = 0.5$ as a function of predictor-predictor correlation b; (**B**) PID with I_{ccs} for $a = c = 0.5$; (**C**) PID with I_{mmi} for $a = 0.4, c = 0.6$; (**D**) PID with I_{ccs} for $a = 0.4, c = 0.6$.

8. Discussion

We have presented I_{ccs}, a novel measure of redundant information based on the expected pointwise change in surprisal that is common to all input sources. Developing a meaningful quantification of redundant and synergistic information has proved challenging, with disagreement about even the basic axioms and properties such a measure should satisfy. Therefore, here we take a bottom-up approach, starting by defining what we think redundancy should measure at the pointwise level (common change in surprisal), and then exploring the consequences of this through a range of examples.

This new redundancy measure has several advantages over existing proposals. It is conceptually simple: it measures precisely the pointwise contributions to the mutual information which are shared unambiguously among the considered sources. This seems a close match to an intuitive definition of redundant information. I_{ccs} exploits the additivity of surprisal to directly measure the pointwise overlap as a set intersection, while removing the ambiguities that arise due to the conflation of pointwise information and misinformation effects by considering only terms with common sign (since a common sign is a prerequisite for there to be a common change in surprisal). I_{ccs} is defined for any number of input sources (implemented for 2 and 3 predictor systems), as well as any continuous system (implemented for multivariate Gaussian predictors and targets). Matlab code implementing the measure accompanies this article [47]. The code requires installation of Python and the `dit` toolbox [50]. The repository includes all the examples described herein, and it is straightforward for users to apply the method to any other systems or examples they would like.

To motivate the choice of joint distribution we use to calculate I_{ccs} we review and extend the decision theoretic operational argument of Bertschinger et al. [12]. We show how a game theoretic operationalisation provides a different perspective, and give a specific example where an exploitable game-theoretic advantage exists for each agent, but I_{broja} suggests there should be no unique information. We therefore conclude the decision theoretic formulation is too restrictive and that the balance of unique and redundant information is not invariant to changes in the predictor-predictor marginal distribution. This means that the optimisation in I_{broja} is not only minimising synergy, but could actually be increasing redundancy. Detailed consideration of several examples shows that the I_{broja} optimisation often results in distributions with coupled predictor variables, which maximises the source redundancy between them. For example, in the SUM system, the coupled predictors make the $(0,0,0)$ and $(1,1,2)$ events redundant, when in the true system the predictors are independent, so those events contribute unique information. However, we note that if required I_{ccs} can also be calculated following the decision theoretic perspective simply by using $\hat{P}_{\text{ind}}$.

I_{ccs} satisfies most of the core axioms for a redundancy measure, namely symmetry, self-redundancy and a modified identity property which reflects the fact that mutual information can itself include synergistic entropy effects [36]. Crucially, it also satisfies subset equality which has not previously been considered separately from monotonicity, but is the key axiom which allows the use of the reduced redundancy lattice. However, we have shown that I_{ccs} is not monotonic on the redundancy lattice because nodes can convey unique misinformation. This means the resulting PID is not non-negative. In fact, negative terms can occur even without non-monotonicity because for some systems (e.g., 3 predictor systems with XOR structures) mechanistic redundancy can result in negative terms [36]. We argue that while "negative … atoms can subjectively be seen as flaw" [37] in fact, they are a necessary consequence of a redundancy measure that genuinely quantifies overlapping information content. We have shown that despite the negative values, I_{ccs} provides intuitive and consistent PIDs across a range of example systems drawn from the literature.

Mutual information itself is an expectation over positive and negative terms. While Jensen's inequality ensures that the overall expectation is non-negative, we argue there is no way to apply Jensen's inequality to decomposed partial information components of mutual information, whichever redundancy measure is used, and thus no reason to assume they must be non-negative. An alternative way to think about the negative values is to consider the positive and negative contributions to mutual information separately. The definition of I_{ccs} could easily be expanded to quantify redundant pointwise information separately from redundant pointwise misinformation (rows 1 and 3 of Table 3). One could then imagine two separate lattice decompositions, one for the pointwise information (positive terms) and one for the pointwise misinformation (negative terms). We conjecture that both of these lattices would be monotonic, and that the non-monotonicity of the I_{ccs} PID arises as a net effect from taking the difference between these. This suggests it may be possible to obtain zero unique information from a cancellation of redundant information with redundant misinformation, analogous to how zero co-information can result in the presence of balanced redundant and synergistic effects, and so exploring this approach is an interesting area for future work. It is also important to develop more formal analytical results proving further properties of the measure, and separate local information versus local misinformation lattices might help with this.

Rauh [55] recently explored an interesting link between the PID framework and the problem of cryptographic secret sharing. Intuitively, there should be a direct relationship between the two notions: an authorized set should have only synergistic information about the secret when all elements of the set are considered, and a shared secret scheme corresponds to redundant information about the secret between the authorized sets. Therefore, any shared secret scheme should yield a PID with a single non-negative partial information term equal to the entropy of the secret at the node representing the redundancy between the synergistic combinations of each authorised set within the inclusion-minimal access structure. Rauh [55] shows that if this intuitive relationship holds, then the PID cannot be non-negative. This finding further supports our suggestion that it may not be possible to obtain a non-negative PID from a redundancy measure that meaningfully quantifies overlapping information content; if such a measure satisfies the intuitive "secret sharing property" [55] it does not provide a non-negative PID. We note that I_{ccs} satisfies the secret sharing property for ([55], Example 1); whether it can be proved to do so in general is an interesting question for future research. These considerations suggest I_{ccs} might be useful in cryptographic applications.

Another important consideration for future research is how to address the practical problems of limited sampling bias [56] when estimating PID quantities from experimental data. Similarly, how best to perform statistical inference with non-parametric permutation methods is an open question. We suggest it is likely that different permutation schemes might be needed for the different PID terms, since trivariate conditional mutual information requires a different permutation scheme than bivariate joint mutual information [57].

How best to practically apply the PID to systems with more than three variables is also an important area for future research. The four variable redundancy lattice has 166 nodes, which already presents a significant challenge for interpretation if there are more than a handful of non-zero partial information values. We suggest that it might be useful to collapse together the sets of terms that have the same order structure. For example, for the three variable lattice the terms within the layers could be represented as shown in Table 14. While this obviously does not give the complete picture provided by the full PID, it gives considerably more detail than existing measures based on maximum entropy subject to different order marginal constraints, such as connected information [43]. We hope it might provide a more tractable practical tool that can still give important insight into the structure of interactions for systems with four or more variables.

Table 14. Order-structure terms for the three variable lattice. Resulting values for the example systems of a giant bit, even parity and DBLXOR (Section 6) are shown.

Level	Order-Structure Terms	Giant Bit	Parity	DBLXOR
7	(3)	0	1	-1
6	(2)	0	0	3
5	(2,2)	0	0	0
4	(1), (2,2,2)	0, 0	0, 0	0, 0
3	(1,2)	0	0	0
2	(1,1)	0	0	0
1	(1,1,1)	1	0	0

We have recently suggested that the concepts of redundancy and synergy apply just as naturally to entropy as to mutual information [36]. Therefore, the redundancy lattice and PID framework can be applied to entropy to obtain a partial entropy decomposition. A particular advantage of the entropy approach is that it provides a way to separately quantify source and mechanistic redundancy [11,36]. Just as mutual information is derived from differences in entropies, we suggest that partial information terms should be related to partial entropy terms. For any partial information decomposition, there should be a compatible partial entropy decomposition. We note that I_{ccs} is highly consistent with a PID based on a partial entropy decomposition obtained with a pointwise entropy redundancy measure which measures common surprisal [36]. More formal study of the relationships between the two approaches is an important area for future work. In contrast, it is hard to imagine an entropy decomposition compatible with I_{broja}. In fact, we have shown that I_{broja} is fundamentally incompatible with the notion of synergistic entropy. Since it satisfies the Harder et al. identity axiom, it induces a two variable entropy decomposition which always has zero synergistic entropy.

As well as providing the foundation for the PID, a conceptually well-founded and practically accessible measure of redundancy is a useful statistical tool in its own right. Even in the relatively simple case of two experimental dependent variables, a rigorous measure of redundancy can provide insights about the system that would not be possible to obtain with classical statistics. The presence of high redundancy could indicate a common mechanism is responsible for both sets of observations, whereas independence would suggest different mechanisms. To our knowledge the only established approaches that attempt to address such questions in practice are Representational Similarity Analysis [58] and cross-decoding methods such as the temporal generalisation method [59]. However, both these approaches can be complicated to implement, have restricted domains of applicability and cannot address synergistic interactions. We hope the methods presented here will provide a useful and accessible alternative allowing statistical analyses that provide novel interpretations across a range of fields.

Acknowledgments: I thank Jim Kay, Michael Wibral, Ryan James, Johannes Rauh and Joseph Lizier for useful discussions. I thank Ryan James for producing and maintaining the excellent dit package. I thank Daniel Chicharro for introducing me to the topic and providing many patient explanations and examples. I thank Eugenio Piasini, Christoph Daube and Philippe Schyns for useful comments on the manuscript. This work was supported by the Wellcome Trust [107802/Z/15/Z] and the Multidisciplinary University Research Initiative/Engineering and Physical Sciences Research Council [EP/N019261/1].

Conflicts of Interest: The author declares no conflict of interest.

References

1. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
2. Cover, T.; Thomas, J. *Elements of Information Theory*; Wiley: New York, NY, USA, 1991.
3. Ince, R.A.; Giordano, B.L.; Kayser, C.; Rousselet, G.A.; Gross, J.; Schyns, P.G. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Hum. Brain Mapp.* **2017**, *38*, 1541–1573.
4. Sokal, R.R.; Rohlf, F.J. *Biometry*; WH Freeman and Company: New York, NY, USA, 1981.
5. Timme, N.; Alford, W.; Flecker, B.; Beggs, J.M. Synergy, redundancy, and multivariate information measures: An experimentalist's perspective. *J. Comput. Neurosci.* **2013**, *36*, 119–140.
6. Williams, P.L.; Beer, R.D. Nonnegative Decomposition of Multivariate Information. *Physics* **2010**, *1004*, 2515.
7. Wibral, M.; Priesemann, V.; Kay, J.W.; Lizier, J.T.; Phillips, W.A. Partial information decomposition as a unified approach to the specification of neural goal functions. *Brain Cogn.* **2017**, *112*, 25–38.
8. Lizier, J.T.; Prokopenko, M.; Zomaya, A.Y. A Framework for the Local Information Dynamics of Distributed Computation in Complex Systems. In *Guided Self-Organization: Inception*; Prokopenko, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 115–158, doi:10.1007/978-3-642-53734-9_5.
9. Reza, F.M. *An Introduction to Information Theory*; McGraw-Hill: New York, NY, USA, 1961.
10. Griffith, V.; Koch, C. Quantifying Synergistic Mutual Information. In *Guided Self-Organization: Inception*; Prokopenko, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 159–190.
11. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. *Phys. Rev.* **2013**, *87*, 012130.
12. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying Unique Information. *Entropy* **2014**, *16*, 2161–2183.
13. Griffith, V.; Chong, E.K.P.; James, R.G.; Ellison, C.J.; Crutchfield, J.P. Intersection Information Based on Common Randomness. *Entropy* **2014**, *16*, 1985–2000.
14. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J. Shared Information—New Insights and Problems in Decomposing Information in Complex Systems. In *Proceedings of the European Conference on Complex Systems 2012*; Gilbert, T., Kirkilionis, M., Nicolis, G., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2013; pp. 251–269, doi:10.1007/978-3-319-00395-5_35.
15. Olbrich, E.; Bertschinger, N.; Rauh, J. Information Decomposition and Synergy. *Entropy* **2015**, *17*, 3501–3517.
16. Griffith, V.; Ho, T. Quantifying Redundant Information in Predicting a Target Random Variable. *Entropy* **2015**, *17*, 4644–4653.
17. McGill, W.J. Multivariate information transmission. *Psychometrika* **1954**, *19*, 97–116.
18. Jakulin, A.; Bratko, I. Quantifying and Visualizing Attribute Interactions. *arXiv* **2003**, arXiv:cs/0308002.
19. Bell, A.J. The co-information lattice. In Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japan, 1–4 April 2003; pp. 921–926.
20. Matsuda, H. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Phys. Rev.* **2000**, *62*, 3096–3102.

21. Wibral, M.; Lizier, J.; Vögler, S.; Priesemann, V.; Galuske, R. Local active information storage as a tool to understand distributed neural information processing. *Front. Neuroinf.* **2014**, *8*, 1.
22. Lizier, J.T.; Prokopenko, M.; Zomaya, A.Y. Local information transfer as a spatiotemporal filter for complex systems. *Phys. Rev.* **2008**, *77*, 026110.
23. Wibral, M.; Lizier, J.T.; Priesemann, V. Bits from Biology for Computational Intelligence. *Quant. Biol.* **2014**, *185*, 1115–1117.
24. Van de Cruys, T. Two Multivariate Generalizations of Pointwise Mutual Information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 16–20.
25. Church, K.W.; Hanks, P. Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.* **1990**, *16*, 22–29.
26. Barrett, A.B. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Phys. Rev.* **2015**, *91*, 052802.
27. Han, T.S. Multiple mutual informations and multiple interactions in frequency data. *Inf. Control* **1980**, *46*, 26–45.
28. Gawne, T.; Richmond, B. How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.* **1993**, *13*, 2758–2771.
29. Panzeri, S.; Schultz, S.; Treves, A.; Rolls, E. Correlations and the encoding of information in the nervous system. *Proc. Biol. Sci.* **1999**, *266*, 1001–1012.
30. Brenner, N.; Strong, S.; Koberle, R.; Bialek, W.; Steveninck, R. Synergy in a neural code. *Neural Comput.* **2000**, *12*, 1531–1552.
31. Schneidman, E.; Bialek, W.; Berry, M. Synergy, Redundancy, and Independence in Population Codes. *J. Neurosci.* **2003**, *23*, 11539–11553.
32. Ting, H. On the Amount of Information. *Theory Prob. Appl.* **1962**, *7*, 439–447.
33. Quian Quiroga, R.; Panzeri, S. Extracting information from neuronal populations: Information theory and decoding approaches. *Nat. Rev. Neurosci.* **2009**, *10*, 173–185.
34. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics: Berlin/Heidelberg, Germany, 2001; Volume 1.
35. Crampton, J.; Loizou, G. The completion of a poset in a lattice of antichains. *Int. Math. J.* **2001**, *1*, 223–238.
36. Ince, R.A.A. The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. *arXiv* **2017**, arXiv:1702.01591.
37. James, R.G.; Crutchfield, J.P. Multivariate Dependence Beyond Shannon Information. *arXiv* **2016**, arXiv:1609.01233.
38. DeWeese, M.R.; Meister, M. How to measure the information gained from one symbol. *Netw. Comput. Neural Syst.* **1999**, *10*, 325–340.
39. Butts, D.A. How much information is associated with a particular stimulus? *Netw. Comput. Neural Syst.* **2003**, *14*, 177–187.
40. Osborne, M.J.; Rubinstein, A. *A Course in Game Theory*; MIT Press: Cambridge, MA, USA, 1994.
41. Jaynes, E. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
42. Amari, S. Information Geometry of Multiple Spike Trains. In *Analysis of Parallel Spike Trains*; Grün, S., Rotter, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 221–252.
43. Schneidman, E.; Still, S.; Berry, M., II; Bialek, W. Network Information and Connected Correlations. *Phys. Rev. Lett.* **2003**, *91*, 238701.
44. Ince, R.; Montani, F.; Arabzadeh, E.; Diamond, M.; Panzeri, S. On the presence of high-order interactions among somatosensory neurons and their effect on information transmission. *J. Phys. Conf. Ser.* **2009**, *197*, 012013.
45. Roudi, Y.; Nirenberg, S.; Latham, P. Pairwise Maximum Entropy Models for Studying Large Biological Systems: When They Can Work and When They Can't. *PLoS Comput. Biol.* **2009**, *5*, e1000380.
46. Lizier, J.T.; Flecker, B.; Williams, P.L. Towards a synergy-based approach to measuring information modification. In Proceedings of the 2013 IEEE Symposium on Artificial Life (ALIFE), Singapore, 16–19 April 2013; pp. 43–51.
47. Robince/partial-info-decomp. Available online: https://github.com/robince/partial-info-decomp (accessed on 29 June 2017).
48. Dit. Available online: https://github.com/dit/dit (accessed on 29 June 2017).

49. Dit: Discrete Information Theory. Available online: http://docs.dit.io/ (accessed on 29 June 2017).
50. James, R.G. cheebee7i. Zenodo. dit/dit v1.0.0.dev0 [Data set]. Available online: https://zenodo.org/record/235071#.WVMJ9nuVmpo (accessed on 28 June 2017).
51. Kay, J.W. On finding trivariate binary distributions given bivariate marginal distributions. Personal Communication, 2017.
52. Abdallah, S.A.; Plumbley, M.D. A measure of statistical complexity based on predictive information with application to finite spin systems. *Phys. Lett.* **2012**, *376*, 275–281.
53. Rauh, J.; Bertschinger, N.; Olbrich, E.; Jost, J. Reconsidering unique information: Towards a multivariate information decomposition. In Proceedings of the 2014 IEEE International Symposium on Information Theory (ISIT), Honolulu, HI, USA, 29 June–4 July 2014.
54. Chicharro, D.; Panzeri, S. Synergy and Redundancy in Dual Decompositions of Mutual Information Gain and Information Loss. *Entropy* **2017**, *19*, 71.
55. Rauh, J. Secret Sharing and Shared Information. *arXiv* **2017**, arXiv:1706.06998.
56. Panzeri, S.; Senatore, R.; Montemurro, M.A.; Petersen, R.S. Correcting for the Sampling Bias Problem in Spike Train Information Measures. *J. Neurophys.* **2007**, *96*, 1064–1072.
57. Ince, R.A.A.; Mazzoni, A.; Bartels, A.; Logothetis, N.K.; Panzeri, S. A novel test to determine the significance of neural selectivity to single and multiple potentially correlated stimulus features. *J. Neurosci. Methods* **2012**, *210*, 49–65.
58. Kriegeskorte, N.; Mur, M.; Bandettini, P. Representational Similarity Analysis—Connecting the Branches of Systems Neuroscience. *Front. Syst. Neurosci.* **2008**, *2*, 4, doi:10.3389/neuro.06.004.2008.
59. King, J.R.; Dehaene, S. Characterizing the dynamics of mental representations: The temporal generalization method. *Trends Cogn. Sci.* **2014**, *18*, 203–210.

© 2017 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).