Washington University in St. Louis
Washington University Open Scholarship

McKelvey School of Engineering Theses & Dissertations | McKelvey School of Engineering

Spring 5-11-2024

Capturing Higher-order Relationships through Information Decomposition

Aobo Lyu
Washington University – McKelvey School of Engineering

Follow this and additional works at: https://openscholarship.wustl.edu/eng_etds

 Part of the Data Science Commons, Probability Commons, and the Theory and Algorithms Commons

Recommended Citation

Lyu, Aobo, "Capturing Higher-order Relationships through Information Decomposition" (2024). *McKelvey School of Engineering Theses & Dissertations*. 1015.
https://openscholarship.wustl.edu/eng_etds/1015

This Thesis is brought to you for free and open access by the McKelvey School of Engineering at Washington University Open Scholarship. It has been accepted for inclusion in McKelvey School of Engineering Theses & Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS

McKelvey School of Engineering
Department of Electrical & Systems Engineering

Thesis Examination Committee:
Andrew Clark, Chair
Xudong Chen
Netanel Raviv

Capturing Higher-order Relationships through Information Decomposition

by
Aobo Lyu

A thesis presented to
the McKelvey School of Engineering
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Science

May 2024
St. Louis, Missouri

© 2024, Aobo Lyu

Table of Contents

List of Figures

Acknowledgments

I would like to extend my deepest gratitude to my supervisor, Professor Andrew Clark, whose detailed guidance over the past two years has profoundly shaped my understanding of rigorous academic research. His mentorship not only introduced me to my ideal scientific research work but also led to highly satisfactory results. Under his stewardship, I mastered the art of completing my tasks with a rhythmic precision and embraced the challenges of academic inquiry.

I am equally thankful to Professor Netanel Raviv, who, despite not being my official supervisor, provided meticulous guidance on every sentence of the Explicit Formula for Partial Information Decomposition project. His thorough revisions and feedback were invaluable, allowing me to approach theoretical research with a newfound calmness.

My appreciation also goes to my committee members, Professor Xudong Chen, whose insightful comments and suggestions significantly enriched my paper and defense preparation.

I am fortunate to have been supported by an incredible circle of friends and colleagues in the research group. Their camaraderie made my journey through these two years not only educational but also immensely enjoyable.

Lastly, I would like to acknowledge the Department of Electrical & Systems Engineering at McKelvey School of Engineering. The department's policies on research and the comprehensive educational opportunities it provided were instrumental in my personal and professional growth. I am immensely grateful for the transformative experiences afforded to me during my time here, which have opened up limitless possibilities for my future.

This work was supported by AFOSR grants FA9550-22-1-0054 and FA9550-23-1-0208.

Aobo Lyu

Washington University in St. Louis
May 2024

ABSTRACT OF THE THESIS

Capturing Higher-order Relationships through Information Decomposition

by

Aobo Lyu

Master of Science in Systems Science & Mathematics

Washington University in St. Louis, 2024

Professor Andrew Clark, Chair

Mutual information between two random variables is a well-studied notion, whose understanding is fairly complete. Mutual information between one random variable and a pair of other random variables, however, is a far more involved notion. Specifically, Shannon's mutual information does not capture fine-grained interactions between those three variables, resulting in limited insights in complex systems. To capture these fine-grained higher-order interactions among variables, Williams and Beer proposed a framework called Partial Information Decomposition (PID) to decompose this mutual information to *information atoms*, called unique, redundant, and synergistic, and proposed several operational axioms that these atoms must satisfy. This conceptual framework provides a potential data-driven approach to reveal higher-order relationships between multiple source variables and a target variable, but still faces many problems, such as incomplete numerical calculations and restricted decomposition scales (multivariable mutual information). In this report, we introduce two works completed in the past semesters, in which one solved numerical calculations problem and another further expanded it to the system scale (whole entropy). In this way, we have the opportunity to implement data-driven methods to reveal higher-order interactions.

The first work is an explicit formula for Partial Information Decomposition. In spite of numerous efforts, a general formula that satisfies all the axioms of PID has yet to be found.

Inspired by Judea Pearl's do-calculus, we resolve this open problem by introducing the *do-operation*, an operation over the variable system which sets a certain marginal to a desired value, which is distinct from any existing approaches. Using this operation, we provide the first explicit formula for calculating the information atoms so that Williams and Beer's axioms are satisfied, as well as additional properties from subsequent studies in the field.

The second work is a framework called System Information Decomposition. Diverging from the PID framework, which concentrates on the directional interactions from an array of source variables to a single target variable, we introduce a novel framework termed System Information Decomposition (SID). By proving all the information atoms are symmetric, the framework can further decompose the whole entropy of the system to capture all interactions among variables. This positions SID as a promising framework with the potential to foster a deeper understanding of higher-order relationships within complex systems across disciplines.

Chapter 1

Introduction

Since its inception by Claude Shannon [31], mutual information has remained a pivotal measure in information theory, which finds extensive applications across multiple other domains. Extending mutual information to multivariate systems has attracted significant academic interest, but no widely agreed upon generalization exists to date. For instance, the so-called *interaction information* [35] emerged in 1960 as an equivalent notion for mutual information in multivariate systems, and yet, it provides negative values in many common systems, contradicting Shannon's viewpoint of information measures as nonnegative quantities.

Arguably the simplest multivariate setting in which Shannon's mutual information fails to capture the full complexity of the system is that of a three variable system, with *two source variables*, and *one target variable*. Mutual information between the source variables and the target variables does not provide insights about *how* the source variables influence the target variable. Specifically, in various points of the probability space the value of the target variable might be computable either:

(a) exclusively from one source variable (but not the other);

(b) either one of the source variables; or

(c) both variables jointly (but not separately).

In 2010 William and Beer [37] proposed to formalize the above fine-grained interactions in a three variable system using an axiomatic approach they called *Partial Information Decomposition* (PID). They proposed decomposing said mutual information to four constituent ingredients called information atoms, which capture the above possible interactions between the variables:

(a) two *unique information atoms*, one for each source variable, which capture the information each source variable implies about the target variable, that cannot be inferred from the other; (b) one *redundant information atom*, which captures the information that can inferred about the target variable from either one of the source variables; and (c) one *synergistic information atom*, which captures the information that can be inferred about the target variable from both source variables jointly, but not individually.

Ref. [37] proposed a set of axioms that the above information atoms should satisfy in order to provide said insights, and follow-up works in the field identified several additional properties [17, 23, 22, 33]. Yet, in spite of extensive efforts [10, 16, 3, 14, 4], a comprehensive definition of information atoms which satisfies all these axioms and properties is yet to be found.

In spite of limited understanding of the information atoms, PID has already found multiple applications in various fields. As a simple example [1, Fig. 2.1], one can imagine the two source variables being education level and gender, and the target variable being annual income. An exact formula for computing the information atoms would shed insightful information about the extent to which annual income is a result of education level, gender, either one, or both.

Beyond this simple example, PID has broad applications in a wide range of fields. In brain network analysis, PID (or similar ideas) has been instrumental in measuring correlations between neurons [30] and understanding complex neuronal interactions in cognitive processes [34]. For privacy and fairness studies, the synergistic concept provides insights about data disclosure mechanisms [27, 13]. In the field of causality, information decomposition can be used to distinguish and quantify the occurrence of causal emergence [29], and more.

In Chapter 3, we introduce the work of **Explicit Formula for Partial Information Decomposition** that satisfies all of Williams and Beer's axioms, as well as several additional desired properties. We do so by introducing the *do-operation*, which is inspired by similar concepts in the field of causal analysis [25, 26, 15]. Intuitively, based on the understanding that unique information is "ideal conditional mutual information," our method first adjusts the entire probability distribution by using the do-operation in order to make the target variable identical to its conditional distribution given one source variable(s), and then calculates the expectation of mutual information between it and the other source variable(s) under different conditions. And it is worth noting that our method is not based on any of the point-wise, localized, or optimization approaches that existing methods use.

In Chapter 4, we introduce the work of **System Information Decomposition**, an innovative theoretically extended framework based on PID that treats all system variables equally (target-free) and effectively captures their complex interactions. Specifically, we firstly expand the PID's conceptual framework to a system horizon by taking all variables in the system as target variable separately. Then, since PID is inspired by an analogy between information theory and set theory [36] and redundant information can be understood as the intersection of variable information, we prove the symmetry properties of information decomposition based on a set theory perspective of information theory. That means the value of information atoms, the non-overlapping units obtained by decomposing variables' information entropy according to their relationship, will not be affected by the the choice of target variable. Therefore, we put forward a general SID framework, wherein redundant, synergistic, and unique information atoms become a multivariate system's property, reflecting the complex (pairwise and higher-order) relationships among variables. Finally, we discuss the potential application scenarios and implications of SID from areas such as Higher-order Networks and theory of Causality.

In Chapter 5, we briefly conclude the rotation result and discuss about and the future works. And all proofs are provided in the appendix.

Chapter 2

Partial Information Decomposition Framework, Axioms, and Properties

The following notational conventions are observed throughout this article: $X, \mathcal{X}, x$ (similarly $Y, \mathcal{Y}, y$ etc.) denote a random variable, its corresponding (finite) alphabet, and an element of that alphabet, respectively. The distribution of X is denoted by $\mathcal{D}_X$, the joint distribution of X and Y is denoted by $\mathcal{D}_{X,Y}$, and the distribution of X given $Y = y$ is denoted by $\mathcal{D}_{X|Y=y}$.

For random variables X, Y, Z, the quantity $I(X, Y; Z)$ captures the amount of information that one *target variable* Z shares with the *source variables* (X, Y), but provides no further information regarding finer interactions between the three variables. To gain more subtle insights into the interactions between Z and (X, Y), [37] proposed to further decompose $I(X, Y; Z)$ into *information atoms*. Specifically, the shared information between Z and (X, Y) should contain a *redundant* information atom, two *unique* information atoms, and one *synergistic* information atom (see Figure 2.1).

The redundant information atom $\mathrm{Red}(X, Y \to Z)$ (also called "shared") represents the information which either X or Y imply about Z. The unique information atom $\mathrm{Un}(X \to Z|Y)$ represents the information individually contributed to Z by X, but not by Y (similarly $\mathrm{Un}(Y \to Z|X)$). The synergistic information atom $\mathrm{Syn}(X, Y \to Z)$ (also called "complementary"), represents the information that can only be known about Z through the *joint* observation of X and Y, but cannot be provided by either one of them separately. Together, we must have that

$$I(X, Y; Z) = \mathrm{Red}(X, Y \to Z) + \mathrm{Syn}(X, Y \to Z) + \mathrm{Un}(X \to Z|Y) + \mathrm{Un}(Y \to Z|X). \quad (2.1)$$

We refer to (2.1) as *Partial Information Decomposition* (PID).

Moreover, since the redundant atom together with one of the unique atoms constitute all information that one source variable implies about the target variable, it must be the case that their summation equals the mutual information between the two, i.e.,

$$\begin{aligned} I(X;Z) &= \mathrm{Red}(X, Y \to Z) + \mathrm{Un}(X \to Z|Y), \\ \text{and } I(Y;Z) &= \mathrm{Red}(X, Y \to Z) + \mathrm{Un}(Y \to Z|X). \end{aligned} \tag{2.2}$$

In a similar spirit, the synergistic information atom and one of the unique information atoms measure shared information between the target variable and one of the source variables, while excluding the other source variable. Therefore, the summation of these quantities should coincide with the well-known definition of conditional mutual information.

$$\begin{aligned} I(Z;X|Y) &= \mathrm{Syn}(X, Y \to Z) + \mathrm{Un}(X \to Z|Y), \\ \text{and } I(Z;Y|X) &= \mathrm{Syn}(X, Y \to Z) + \mathrm{Un}(Y \to Z|X). \end{aligned} \tag{2.3}$$

Eqs. (2.2), and (2.3) are the foundation of an axiomatic approach towards an operational definition of the information atoms. These equations form the first in a series of axioms, presented next, which were raised in previous works on the topic [37, 36, 12]. Such axiomatic approach was also taken in the past in order to shed light on Shannon's mutual information [8].

Axiom 1 (Information atoms relationship)**.** *Partial Information Decomposition* (2.1) *satisfies* (2.2) *and* (2.3)*.*

Notice that it suffices to specify the definition of any one of the information atoms, and the definitions for the remaining atoms follow from Axiom 1. Consequently, [37, 20] chose to specify Red, and provided three additional axioms which Red should satisfy.

The first additional axiom is *commutativity* of the source variables, which implies that the order of the source variables must not affect the value of the redundant information.

Axiom 2 (Commutativity)**.** *Partial Information Decomposition satisfies* $\mathrm{Red}(X, Y \to Z) = \mathrm{Red}(Y, X \to Z)$*.*

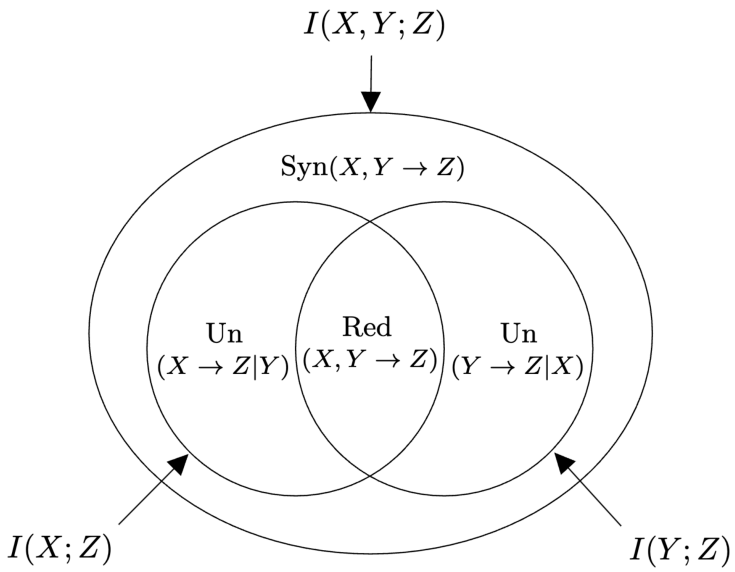

Figure 2.1: A pictorial representation of Partial Information Decomposition (2.1), where $I(X,Y;Z)$ is decomposed to its finer information atoms, the synergistic $\mathrm{Syn}(X,Y \to Z)$ (also called "complementary"), the redundant $\mathrm{Red}(X,Y \to Z)$ (also called "shared"), and the two directional unique components $\mathrm{Un}(X \to Z|Y)$ and $\mathrm{Un}(Y \to Z|X)$. The summation of the redundant atom and one of the unique atoms must be equal to the corresponding mutual information, as described in Eq. (2.2).

The second is *monotonicity*, which implies that the redundant information is non-increasing when adding a source variable, since the newly added variable cannot increase the redundancy between the original variables. We sidestep the discussion about monotonicity with more than two variables, which is not our focus in this paper, even though it can be easily obtained by extending our definition to more than two source variables.

The third is *self-redundancy*, which defines the redundant information from *one* source variable to the target variable (i.e., $\mathrm{Red}(X \to Z)$) as the mutual information between them. In the case of two source variables considered herein, monotonicity and self-redundancy merge into the following single axiom.

Axiom 3 (Monotonicity and self-redundancy)**.** *Partial Information Decomposition satisfies* $\mathrm{Red}(X, Y \to Z) \leq \min\{I(X;Z), I(Y;Z)\}$.

Notice that Axiom 3, alongside Axiom 1 (specifically (2.2)), imply that Un is a nonnegative quantity. The nonnegativity of Red is stated in [37, 20] as a separate axiom, shown next.

Axiom 4 (Nonnegativity)**.** *Partial Information Decomposition satisfies* $\mathrm{Red}(X, Y \to Z) \geq 0$.

The nonnegativity of Syn is normally not listed as an axiom, since it is debatable if it should or should not be nonnegative; we will show that our method yields nonnegative Syn under the *closed-system assumption* (i.e., $H(Z|X,Y) = 0$) in Chapter 3, and further discussion is given in Chapter 3.3.

Besides, subsequent to [37, 20], studies suggested two additional properties, *additivity* and *continuity* [4, 28]. Additivity implies that whenever independent variable systems are considered, the joint information measures should be the sum of the information measures of each individual system. This is the case, for instance, in joint entropy of two independent variables.

Property 1 (Additivity)**.** *Partial Information Decomposition of two independent systems* $\mathcal{D}_{X_1,X_2,X_3}$ *and* $\mathcal{D}_{A_1,A_2,A_3}$ *satisfy*

$$\begin{aligned}\mathrm{Un}((X_i, A_l) \to (X_j, A_m)|(X_k, A_n)) &= \mathrm{Un}(X_i \to X_j|X_k) + \mathrm{Un}(A_l \to A_m|A_n), \text{ and}\\ \mathrm{F}((X_i, A_l), (X_j, A_m) \to (X_k, A_n)) &= \mathrm{F}(X_i, X_j \to X_k) + \mathrm{F}(A_l, A_m \to A_n),\end{aligned}$$

for every $\mathrm{F} \in \{\mathrm{Red}, \mathrm{Syn}\}$ *and every* $i, j, k, l, m, n \in \{1, 2, 3\}$.

Continuity implies that small changes in the probability distribution lead to small changes in the value of the information measure. It ensures that the measure behaves predictably and is a key property in information theory, particularly for measures like entropy and mutual information.

Property 2 (Continuity)**.** Red, Un*, and* Syn *are continuous functions from the underlying joint distributions* $\mathcal{D}_{X,Y,Z}$ *to* $\mathbb{R}$.

In addition, another well-known property is *independent identity* [16], which asserts that in a system of two independent source variables and a target variable which equals to their joint distribution, the redundant information should be zero.

Property 3 (Independent Identity)**.** *If* $I(X, Y) = 0$ *and* $Z = (X, Y)$*, then* $\mathrm{Red}(X, Y \to Z) = 0$.

We mention that several important properties can be inferred from the above. For example, the non-negativity of Un can be obtained from Axiom 1 and Axiom 3 as mentioned earlier; the commutativity of Syn follows from Axiom 1 and Axiom 2; the difference between (2.2) and (2.3) is often called *consistency* [4], etc.

Finally, we emphasize once again that none of the existing operational definitions of the information atoms satisfy all of the above. A comprehensive list of violations is beyond the page limit of this paper, and yet we briefly mention that Axiom 4 (nonnegativity) is violated by [37, 10, 16, 9] (although some sources do not refer to non-negativity as a requirement); Property 1 (additivity) is violated by all works except [4], [11], [10], and [18] according to [28]; Property 3 (independent identity) is violated by [37]; Property 2 (continuity) is violated by [14, 10], [11], [18], etc.

Chapter 3

Explicit Formula for Partial Information Decomposition

In this chapter, we present our operational definition of Un, from which the definitions of the remaining information atoms follow. Then, we prove that this definition satisfies all the axioms and properties proposed in Chapter 2.

3.1 Definition of Information Atoms

Our definition of unique information Un requires a *do-operation*. This newly defined operation generates a new distribution $\mathcal{D}_{A,B,C}$ with a prescribed marginal from the given distribution $\mathcal{D}_{X,Y,Z}$, and is inspired by Judea Pearl's do-calculus [24] (also [15]). Specifically, we write $\mathcal{D}_{A,B,C} = do(\mathcal{D}_{X,Y,Z}|\mathcal{D}_C)$ to indicate that given a joint distribution $\mathcal{D}_{X,Y,Z}$ and a probability distribution $\mathcal{D}_C$ (over the same alphabet as $\mathcal{Z}$), we construct a new distribution $\mathcal{D}_{A,B,C}$ whose rightmost variable has the same marginal distribution as the input distribution $\mathcal{D}_C$ (and hence the notational choice to represent both by the letter C).

Definition 1 (Do-operation)**.** *Given $\mathcal{D}_{X,Y,Z}$ and $\mathcal{D}_C$ with identical support to $\mathcal{D}_Z$, let $\mathcal{D}_{A,B,C} = do(\mathcal{D}_{X,Y,Z}|\mathcal{D}_C)$ be such that*

$$\Pr(A,B,C=x,y,z) = \begin{cases} 0 & \text{if } \Pr(Z=z)=0, \text{ and} \\ \frac{\Pr(X,Y,Z=x,y,z)\Pr(C=z)}{\Pr(Z=z)} & \text{otherwise,} \end{cases} \tag{3.1}$$

for all $x, y, z \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$.

In Lemma 13, which is given and proved in Appendix A.1, it is shown that $\mathcal{D}_{A,B,C}$ in Definition 1 is well-defined, and that the rightmost marginal of $\mathcal{D}_{A,B,C}$ is identical to $\mathcal{D}_C$.

Therefore, there is no ambiguity in referring to both the input distribution and the rightmost marginal of the output distribution by the same letter C. We now turn to present our definition of Un.

Definition 2 (Unique Information)**.** *For $y \in \mathcal{Y}$ let C_y be a random variable with distribution $\mathcal{D}_{C_y} = \mathcal{D}_{Z|Y=y}$, and let $\mathcal{D}_{A_y,B_y,C_y} = do(\mathcal{D}_{X,Y,Z}|\mathcal{D}_{C_y})$. The unique information from X to Z given Y is defined as:*

$$\mathrm{Un}(X \to Z|Y) = \sum_{y\in\mathcal{Y}} \Pr(Y = y) I(A_y; C_y). \tag{3.2}$$

The definitions for the remaining information atoms are then implied by Axiom 1 as follows.

Definition 3 (Redundant Information)**.** *The Redundant Information from X and Y to Z is defined as:*

$$\mathrm{Red}(X, Y \to Z) = I(X; Z) - \mathrm{Un}(X \to Z|Y).$$

Definition 4 (Synergistic Information)**.** *The synergistic information from X and Y to Z is defined as:*

$$\mathrm{Syn}(X, Y \to Z) = I(X; Z|Y) - \mathrm{Un}(X \to Z|Y).$$

It should be noted that Definition 3 and Definition 4 strictly depend on the order of the source variables; the commutativity of Red (Axiom 2) will be addressed in the sequel, and the commutativity of Syn follows from Axiom 1 and Axiom 2 as mentioned earlier.

3.2 Satisfaction of axioms and properties

To show that our definition satisfies the axioms and properties mentioned in Section 2, we require the following technical lemma, which shows that conditional entropy can be written using the do-operation. The proof is given in Appendix A.2.

Lemma 1. *Following the notations of Definition 2, we have that $H(X|Z) = \sum_{y\in\mathcal{Y}} \Pr(Y = y) H(A_y|C_y)$.*

Moreover, since $I(A_y; C_y) = H(A_y) - H(A_y|C_y)$, by Definition 2 and Lemma 1, the following is immediate.

Corollary 1. *Unique information (Def. 2) can also be written as:*

$$\mathrm{Un}(Y \to Z|X) = \sum_{y \in \mathcal{Y}} \Pr(Y = y) H(A_y) - H(X|Z).$$

We now turn to define a new auxiliary random variable $A_{Z|Y}$, which is defined through its conditioned probabilities.

Definition 5. *For a given (X, Y, Z), let $A_{Z|Y}$ be a random variable over $\mathcal{X}$ with*

$$\Pr(A_{Z|Y} = x|Y = y) = \Pr(A_y = x), \tag{3.3}$$

which implies that

$$\Pr(A_{Z|Y} = x) = \sum_{y \in \mathcal{Y}} \Pr(Y = y) \Pr(A_y = x). \tag{3.4}$$

The fact that $A_{Z|Y}$ is well-defined is proved in Appendix A.3. And by (3.3), we have the following corollary.

Corollary 2. *For every $y \in \mathcal{Y}$, the above A_y and $A_{Z|Y}$ satisfy*

$$H(A_{Z|Y}|Y = y) = H(A_y).$$

Furthermore, the proof in Appendix A.3 also shows the following.

Lemma 2. *The variable $A_{Z|Y}$ above satisfies $H(A_{Z|Y}) = H(X)$.*

So far, we require one final auxiliary lemma, which is based on Corollary 2 and Lemma 2 and proved in Appendix A.4.

Lemma 3. *We have*

$$\sum_{y \in \mathcal{Y}} \Pr(Y = y) H(A_y) \le H(X). \tag{3.5}$$

Based on the above lemmas and corollaries, we are in a position to prove that our definition of Un satisfies the required axioms.

3.2.1 Proof of Axiom 1, Information atoms relationship

Follows immediately from Definition 3 and Definition 4.

3.2.2 Proof of Axiom 3, Monotonicity and self-redundancy

According to Definition 2, Un is nonnegative since it is an expectation of mutual information quantities. Therefore, Axiom 3 follows directly from Definition 3.

3.2.3 Proof of Axiom 4, Nonnegativity

We begin by showing that Red is nonnegative, for which we require the following lemma, proved in Appendix A.5.

Lemma 4. *Unique information (Definition 2) is bounded from above by mutual information, i.e.,*

$$\mathrm{Un}(X \to Z|Y) \leq I(X;Z).$$

Then, nonnegativity of Red (in Def. 3) follows from Lemma 4:

Corollary 3 (Nonnegativity of Redundant Information)**.** *Redundant information (Definition 3) is nonnegative, i.e.,*

$$\mathrm{Red}(X, Y \to Z) \geq 0.$$

In addition to Lemma 4, the unique information can also be proved to be smaller than the conditional entropy, as shown next and proved in Appendix A.6.

Lemma 5. *The unique information defined in Definition 2 is bounded above by conditional information, such that:*

$$\mathrm{Un}(X \to Z|Y) \leq H(Z|Y). \tag{3.6}$$

Although it is not a required property, the nonnegtivity of Syn will follow from Lemma 5 with an additional closed system assumption $H(Z|X,Y) = 0$.

Corollary 4 (Nonnegativity of Synergistic Information)**.** *If $H(Z|X,Y) = 0$, then Synergistic information (Definition 4) is nonnegative, i.e.,* $\mathrm{Syn}(X,Y \to Z) \geq 0$.

3.2.4 Proof of Axiom 2, Commutativity

First, Definition 5 provides an equivalent way to compute Red in the following lemma, which is proved in Appendix A.7.

Lemma 6. *Redundant information (Definition 3) can alternatively be written as* $\mathrm{Red}(X,Y \to Z) = I(A_{Z|Y};Y)$.

Similarly, by switching between X and Y in Definition 3 we have that $\mathrm{Red}(Y,X \to Z) = I(Y;Z) - \mathrm{Un}(Y \to Z|X)$; based on Lemma 6, this equals to $I(X;B_{Z|X})$, where $B_{Z|X}$ is defined analogously to $A_{Z|Y}$ (Definition 5), i.e.,

$$\Pr(B_{Z|X} = y) = \sum_{x \in \mathcal{X}} \Pr(X = x) \Pr(B_x = y).$$

Then, we can conclude the commutativity of redundant information through the following lemma, which is proved in Appendix A.8.

Lemma 7 (Commutativity of Redundant Information)**.** *For $A_{Z|Y}$ and $B_{Z|X}$ as above, we have that $I(A_{Z|Y};Y) = I(X;B_{Z|X})$.*

Combining Lemma 6 and Lemma 7 readily implies the commutativity of Red, i.e., $\mathrm{Red}(X,Y \to Z) = \mathrm{Red}(Y,X \to Z)$.

3.2.5 Proof of Property 1, Additivity

The following lemma is proved in Appendix A.9.

Lemma 8 (Additivity of Unique Information)**.** *For two independent sets of variables X, Y, Z and X', Y', Z', unique information (Definition 2) is additive:*

$$\mathrm{Un}((X, X') \to (Z, Z')|(Y, Y')) = \mathrm{Un}(X \to Z|Y) + \mathrm{Un}(X' \to Z'|Y').$$

Since mutual information and conditional entropy are additive in the above sense, by Definition 3 and Definition 4, alongside Lemma 8, Red and Syn are additive as well.

3.2.6 Proof of Property 2, Continuity

We begin by showing that Red is continuous, for which we require the following lemma proved in Appendix A.10.

Lemma 9 (Continuity of Redundant Information)**.** *The redundant information (Def. 3) is a continuous function of the input distribution $\mathcal{D}_{X,Y,Z}$ to $\mathbb{R}$.*

By Definition 3 and Definition 4, the continuity of Un and Syn can also be derived.

3.2.7 Proof of Property 3, Independent Identity

The following lemma is proved in Appendix A.11.

Lemma 10. *The operator* Red *satisfies Property 3.*

3.3 Discussion about the Formula

In this paper, we proposed an explicit operational formula for PID, which is distinct from any existing approach, and proved that it satisfies all axioms and properties. In this section we provide an intuitive explanation for our approach.

First, we wish to elucidate the role that our do-operation plays in the definition of Un (Definition 2). In a sense, the do-operation can be understood as adjusting the marginal distribution of the Z variable of $\mathcal{D}_{X,Y,Z}$, while impacting its connections with other variables as little as possible. This understanding can be confirmed by Lemma 1, which shows that the expected value of the conditional entropy after the do-operation retains its original value. This resembles the invariance implied in Shannon's communication model [31], where the conditional entropy of the output given the input is not affected by the input distribution. From this perspective, Z and X can be regarded as the input and output of the channel, that indices their "relationship." The do-operation changes the distribution of the input Z, but does not change the channel's characteristic (i.e., $H(Z|X)$).

Based on this, Definition 2 realizes the intuition that unique information should represent the relationship between source variable and target variable given other source variables. So, we use the do-operation to control the marginal distribution of the target variable Z to its conditional distribution given the value y of some source variable(s) Y, then use the expectation of mutual information $\sum_y \Pr(Y = y) I(A_y; C_y)$ to capture the "connection" between the specific source variable X and target variable Z given Y after the do-operation.

The reason this method can partition $I(X; Z|Y)$ to Syn and Un (Def. 4), is that the do-operation eliminates high-order relations between Y and X, Z, i.e. Syn. Specifically, conditional mutual information relies on the joint conditional probability $\mathcal{D}_{(X,Z)|y}$, in expectation over all $y \in \mathcal{Y}$. This distribution includes both the conditional influence of Y on X, Z, but also has a simultaneous influence on the relationship between X and Z.

However, Definition 2 of unique information retains the relationship between X and Z without influence from Y by using the conditional probability $\mathcal{D}_{Z|y}$, in expectation over all $y \in \mathcal{Y}$, to perform the do-operation, which only reflects the conditional influence of Y on Z. Therefore, the expectation of mutual information $\sum_y \Pr(Y = y) I(A_y; C_y)$ can accurately quantify the unique information, which represents the pure conditional mutual relationship.

In addition to the above analysis of do-operations in unique information, Lemma 6 also brings another perspective worth discussing. Redundant information can be understood as the mutual information $I(A_{Z|Y}; Y)$ (or $I(X; B_{Z|X})$) obtained by changing the joint probability distribution $\mathcal{D}_{X,Y}$ according to $\mathcal{D}_{X,Y,Z}$ without changing the marginal distribution $\mathcal{D}_X$ and $\mathcal{D}_Y$ according to Lemma 2 ($H(A_{Z|Y}) = H(X)$).

As mentioned earlier, our definition of Syn might be negative, unless the system is closed (i.e., $H(Z|X,Y) = 0$, Corollary 4). While Un and Red represent the information shared by one or two source variables with the target variable, Syn represents the information provided to the target variable by the "cooperation" of source variables. It is an accepted aphorism that cooperation does not necessarily increase outcome, and hence it might be the case that negative values of Syn conform with intuition. However, the reason why this explanation is no longer necessary in a closed system, as well as alternative interpretations of Syn that are nonnegative, remain to be studied.

Chapter 4

System Information Decomposition

In this Chapter, we develop a mathematical framework called System Information Decomposition. The objective of this framework is to decompose the information of all variables within a system based on their interrelationships. By addressing the limitation of PID, which focuses solely on a single target variable, we progress towards multi-variable information decomposition for systems. Firstly, in Part 4.1, we introduce a perspective on understanding PID to further expand this conceptual model. Afterwards, we simply expanded the decomposition range of PID in Part 4.2. In Part 4.3, by proving the symmetry of information atoms, we obtain the formal form of SID, and the potential applications of this work are discussed in the subsequent Part 4.4.

4.1 Set-theoretic Understanding of PID

Kolchinsky's work [18] offers an understanding based on set theory. Given that PID is inspired by an analogy between information theory and set theory [36], the redundant information can be understood as information sets that the sources provide to the target. More specifically, the definition of set intersection $\cap\{X_i\}$ in set theory means the largest set that is contained in all of the X_i, and these set-theoretic definitions can be mapped into information-theoretic terms by treating "sets" as random variables, "set size" as entropy, and "set inclusion" as an ordering relation $\sqsubset$, which indicates when one random variable is more informative than another.

Considering a set of sources variables $X_1, ..., X_n$ and a target Y, PID aims to decompose $I(X_i, X_j \to Y)$ and get $Red(X_1, \cdots, X_n \to Y)$, the total same information provided by all sources about the target, into a set of non-negative terms. Therefore, redundant information

can be viewed as the "intersection" of the information contributed by different sources, leading to the following definition:

Definition 6 (Set Intersection of Information [18])**.** *For a variable-system, the redundant information from the source variables $X_1, \cdots, X_n$ to the target variable Y is the information that all source variables can provide to the target variable, the largest mutual information between the target variable and a non-unique variable Q that has an ordering relation $\sqsubset$ with all source variables. That is*

$$Red(X_1, \cdots, X_n \to Y) = I_\cap(X_1, \cdots, X_n \to Y) := \sup_Q \{I(Q:Y) : Q \sqsubset X_i, \forall i \in \{1 \cdots n\}\} \tag{4.1}$$

The ordering relation $\sqsubset$ is an analogy to the relation contained $\subseteq$ in set theory, which is not specified but follows some assumptions: i) Monotonicity of mutual information, $A \sqsubset B \Rightarrow I(A:Y) \leq I(B:Y)$. ii) Reflexivity: $A \sqsubset A$ for all variable A. iii) For all sources X_i, $O \sqsubset X_i \sqsubset (X_1, \cdots, X_n)$, where $H(O) = 0$ and $(X_1, \cdots, X_n)$ indicates all sources considered jointly. For example, the partial order can be Q $\sqsubset$ X if and only if $H(Q|X) = 0$, or the well-known Blackwell order [6], such that Q precedes X_i if X_i has all of the information that Q has, about some third variable Y.

4.2 Extension of PID in a System Scenario

The PID method only decomposes joint mutual information between multiple source variables and a specific target variable, as illustrated by the outermost circle of the Venn diagram in Figure 2.1. We redesign the Venn diagram with adding the joint conditional entropy of Y and the conditional entropy of the source variable X_1 and X_2 to obtain a system-wide perspective, as demonstrated in Figure 4.1. The system comprises two source variables,X_1 and X_2, and one target variable, Y, represented by the three intersecting circles.

The area size within the figure signifies the information entropy of the variables or information atoms, and the central area denotes the joint mutual information, encompassing redundant, unique from X_1, unique from X_2, and synergistic information. This arrangement aligns with the Venn diagram framework of PID.

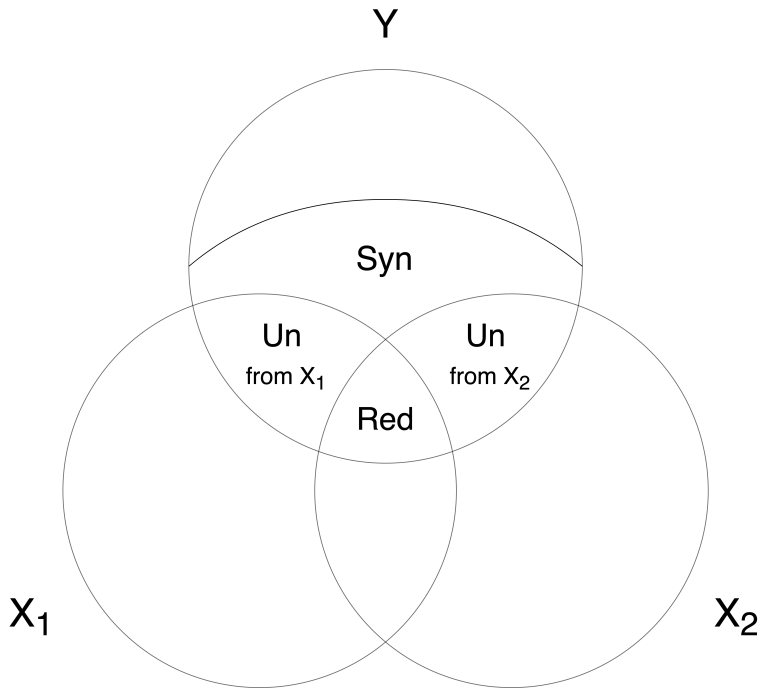

Figure 4.1: Venn diagram from different perspectives of PID.

To enhance the comprehensiveness of the framework, it is necessary to elucidate the unexplored section of the updated Venn diagram 4.1. In addition to the four sections of joint mutual information, the information entropy of the target variable Y contains an unaccounted-for area. According to Shannon's formula, this area corresponds to the joint conditional entropy of the source variables to the target variable $H(Y|X_1, X_2)$, which also characterizes the interrelationships between the target variable and the source variables. In the SID framework, numerous joint conditional entropy exist, including one that stands out: the joint conditional entropy originating from all variables except the target variable. To optimize the usefulness of the SID framework, we define this specific joint conditional entropy as the target variable's external information (Ext). The definition is grounded in the philosophical assumption that everything is interconnected. Since joint conditional entropy implies the uncertainty that cannot be eliminated by the internal variables of the system, the variables capable of providing this information must exist outside the system. To some extent, external information can emphasize the relationship between the target variable and the entire system rather than just a simple relationship with other variables. Therefore, we also consider it a kind of information atom within the SID framework.

Definition 7 (External Information)**.** *For a system containing variables* Y *and* $\{X_1, \cdots, X_n\}$*, the external information* $Ext(Y)$ *is defined as:*

$$Ext(Y) = H(Y|X_1, X_2, \cdots, X_n) \tag{4.2}$$

Thus, we have been able to decompose the target variable's entropy into a finite number of non-repeated information atoms according to the relationship between it and the other variables. Furthermore, we can apply this extended PID framework with three variables as target variable respectively to decompose the entire information entropy of the system, which results in a SID's preliminary version. For the convenience of expression, we use Un_{i-j}, Syn_{ij-k}, and Red_{ij-k} to represent Un, Syn, and Red respectively. A Venn diagram for a three-variable system is shown in Figure 4.2:

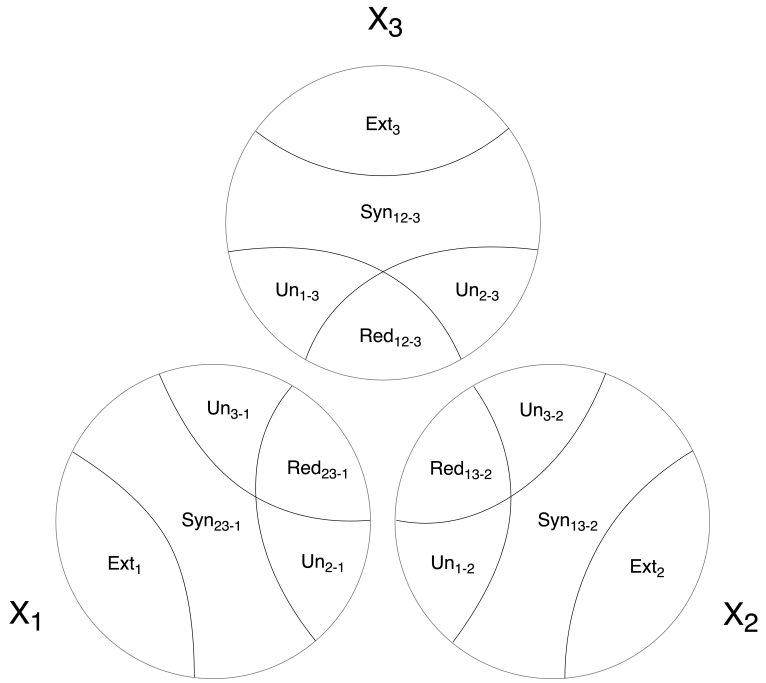

Figure 4.2: Venn diagram of SID's Preliminary version.

4.3 Properties of Information Atoms

Although the preliminary version of SID can decompose all variables in a system, the decomposition of each variable is carried out separately, and the description of information

atoms is directional (from source variables to the target variable). For instance, the unique information provided by X_1 to X_3 in Fig. 4.2 is not directly related to the unique information provided by X_3 to X_1. To make information atoms better reflect the relationship among variables and unifies the Venn diagram of Shannon's framework and the PID framework, it is necessary to further explore the properties of information atoms within the SID framework. In this section, we are going to prove the symmetry property of information atoms by demonstrating that unique, redundant, and synergistic information atoms remain stable when different variables are considered as target variables.

Theorem 1 (Symmetry of Redundant Information)**.** *Let $X_1, \cdots, X_n$ be the variables in a system. The redundant information is equal irrespective of the chosen target variable. Formally, we write $Red(X_i : X_1, \cdots, X_n \setminus X_i) = Red(X_j : X_1, \cdots, X_n \setminus X_j), \forall i, j \in \{1 \cdots n\}$.*

To proof this Theorem, we use the Definition 6 with the Blackwell partial order, such that Q precedes X_i if X_i has all of the information that Q has, about the target variable Y, which written in the form $Q \sqsubset_Y X_i$. The proof is given in Appendix A.12. Then, we have the following lemmas, which is proved in Appendix A.13 and A.14.

Lemma 11 (Symmetry of Unique Information)**.** *Let $X_1, \cdots, X_n$ be the variables in a system. In SID, the unique information of any two variables relative to each other is equal, regardless of which is chosen as the target variable. Formally, we write $Un(X_i : X_j) = Un(X_j : X_i)$, $\forall i \neq j$ where $i, j \in \{1, \cdots, n\}$.*

Lemma 12 (Symmetry of Synergistic Information)**.** *Let $X_1, \cdots, X_n$ be the variables in a system. In SID, the synergistic information of any group of variables is equal, regardless of which is chosen as the target variable. Formally, we write $Syn(X_i : \{X_1, \cdots, X_n\} \setminus X_i) = Syn(X_j : \{X_1, \cdots, X_n\} \setminus X_j), \forall i, j \in \{1 \cdots n\}$.*

Based on the Theorem 1, Lemma 11 12 (the symmetry of information atoms), the SID framework can be merged into the formal version in Figure 4.3. In the formal version of SID, the concept of target variable is canceled, and all variables are equally decomposed according to their relationship with other variables. Specifically, redundant information and unique information are merged. Redundant information (atoms) in any group of variables and unique information (atoms) between any two variables appear only shown one time in the Venn diagram. And each variable contains one external information (atom). While synergistic information (atoms) cannot be fused in a two-dimensional plane, we present them

independently and give them the same symbol Syn_{123} (also value and area). So far, we can give the formal definition of SID:

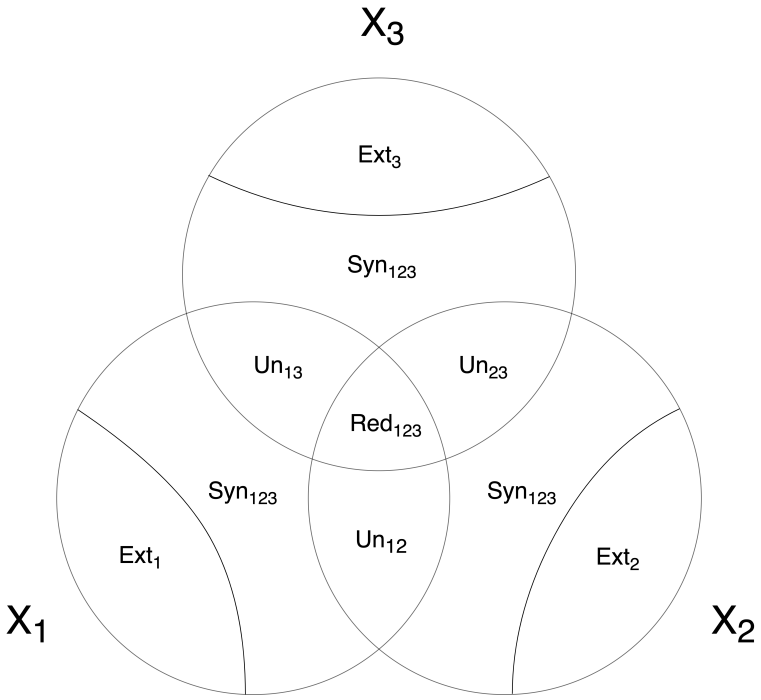

Figure 4.3: Venn diagram of SID's Formal Version.

Definition 8 (System Information Decomposition Framework)**.** *SID is a conceptual system decomposition framework based on information entropy, that can divide the whole information entropy of a multivariate system into non-overlapping information atoms according to the relationship among variables. In this framework, redundant information represents the common or overlapping information of all the variables; unique information represents information that is only owned by two variables but not by others; and synergistic information represents the information that can be known from any variable only when the other variables are observed simultaneously.*

In the SID framework, the Venn diagram unifies the Shannon's framework and PID framework. For a intuitive presentation, we only give the Venn diagram of three-variable system ($\{X_1, X_2, X_3\}$) in this paper.

4.4 Discussion about SID

A foreseeable application across many domains comes from that SID deepens our understanding of data, measures, and information. A worth exploring direction is the quantitative analysis of Higher-order Networks [5]. Since SID can provide a data-driven framework for identifying and analyzing of high-order network structures, it may potentially impact the analysis and understanding of complex systems across various domains [2]. For example, in studying neural networks and brain connectivity [7], the SID framework can provide further insights into the higher-order information flow between multiple neurons or brain regions, which will allow us to directly generate higher-order network models between neurons through the temporal data of multiple neurons, and use this model to explain the implementation of specific functions; in ecological [19], financial, or social systems, the quantitative characterization of high-order relationships among multiple agents can assist in the development of more accurate models and forecasts, as well as the design of effective control methods.

Another field where SID may interact is Causal Science, since it is a field for studying the intrinsic relationships between multiple variables. One of the goals of causal science is to search for invariance in the system. We hope that the revealed properties of the system are independent of the distribution of the data. However, the results obtained from SID can vary with changes in the data distribution. Therefore, adopting the methods of causal science in SID to reveal system invariance is one direction worth to explore.

Apart from the above fields, SID may also has potential applications. Since information atoms provide a more refined division of information entropy, when the physical meaning of information atoms within the SID framework is revealed, specific information atoms may also become indicators for some optimization or learning problems; The symmetry property of synergistic information in SID may provide inspiration for the information disclosure, an important application of PID in information protection field. In summary, SID, as a progress in the underlying measurement, may play a role in many application scenarios, which is also the focus of our next stage of work.

In addition to the above-mentioned promising progress and expectations, there are still some limitations worthy of attention. The first limitation is that the existing proofs of framework properties and computational methods have only been established for three-variable systems.

Although extending current work to general multivariate systems is not a formidable challenge, it contains many aspects of work, such as how to present the decomposition results of multivariate systems on a two-dimensional plane, which will be considered in the next stage of research. In addition, some feasible extensions are also worth exploring, such as make the information decomposition framework into time-resolved approaches [23, 32], and get a point-wise localization [21] are potential topics.

Chapter 5

Conclusion and Future Works

In this work, we complete two pieces of work, such that Explicit Formula for Partial Information Decomposition and System Information Decomposition. The first work provided the first explicit formula for calculating the information atoms so that Williams and Beer's axioms are satisfied, as well as additional properties from subsequent studies in the field. This work solved the most important open problem in the field of Information Decomposition. Also, the System Information Decomposition (SID) framework, by connecting information atoms to higher-order relationships, offers novel insights for decomposing complex systems and analyzing higher-order relationships while addressing the limitations of existing information decomposition methods. In conclusion, the SID framework signifies a promising new direction for investigating complex systems and information decomposition. We anticipate that those works will serve as a valuable tools across an expanding array of fields in the future. In the next stage of research, in addition to the issues mentioned separately in the discussion chapter above, we will integrate the above two works and try to propose a general Higher-order Information Networks Model that can achieve data-driven identification, quantification and reconstruction of Higher-order Interactions (Structures). And try to explore more higher-order networks-related research based on this model.

References

[1] P. K. Banerjee, J. Rauh, and G. Montúfar. Computing the unique information. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 141–145. IEEE, 2018.

[2] F. Battiston, E. Amico, A. Barrat, G. Bianconi, G. Ferraz de Arruda, B. Franceschiello, I. Iacopini, S. Kéfi, V. Latora, Y. Moreno, et al. The physics of higher-order interactions in complex systems. *Nature Physics*, 17(10):1093–1098, 2021.

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

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

[5] G. Bianconi. *Higher-order networks*. Cambridge University Press, 2021.

[6] D. Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, pages 265–272, 1953.

[7] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198, 2009.

[8] I. Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.

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

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

[11] V. Griffith and T. Ho. Quantifying redundant information in predicting a target random variable. *Entropy*, 17(7):4644–4653, 2015.

[12] V. Griffith and C. Koch. Quantifying synergistic mutual information. In *Guided self-organization: inception*, pages 159–190. Springer, 2014.

[13] F. Hamman and S. Dutta. Demystifying local and global fairness trade-offs in federated learning using partial information decomposition. *arXiv preprint arXiv:2307.11333*, 2023.

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

[15] E. P. Hoel, L. Albantakis, and G. Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, 2013.

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

[17] R. A. Ince. The partial entropy decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. *arXiv preprint arXiv:1702.01591*, 2017.

[18] A. Kolchinsky. A novel approach to the partial information decomposition. *Entropy*, 24(3):403, 2022.

[19] S. A. Levin. Self-organization and the emergence of complexity in ecological systems. *Bioscience*, 55(12):1075–1079, 2005.

[20] J. T. Lizier, B. Flecker, and P. L. Williams. Towards a synergy-based approach to measuring information modification. In *2013 IEEE Symposium on Artificial Life (ALIFE)*, pages 43–51. IEEE, 2013.

[21] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. Local measures of information storage in complex distributed computation. *Information Sciences*, 208:39–54, 2012.

[22] A. Lyu, B. Yuan, O. Deng, M. Yang, A. Clark, and J. Zhang. System information decomposition. *arXiv preprint arXiv:2306.08288*, 2023.

[23] P. A. Mediano, F. Rosas, R. L. Carhart-Harris, A. K. Seth, and A. B. Barrett. Beyond integrated information: A taxonomy of information dynamics phenomena. *arXiv preprint arXiv:1909.02297*, 2019.

[24] L. G. Neuberg. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685, 2003.

[25] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[26] J. Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversity-Press*, 19(2):3, 2000.

[27] B. Rassouli, F. E. Rosas, and D. Gündüz. Data disclosure under perfect sample privacy. *IEEE Transactions on Information Forensics and Security*, 15:2012–2025, 2019.

[28] J. Rauh, P. K. Banerjee, E. Olbrich, G. Montúfar, and J. Jost. Continuity and additivity properties of information decompositions. *International Journal of Approximate Reasoning*, 161:108979, 2023.

[29] F. E. Rosas, P. A. Mediano, H. J. Jensen, A. K. Seth, A. B. Barrett, R. L. Carhart-Harris, and D. Bor. Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLoS computational biology*, 16(12):e1008289, 2020.

[30] E. Schneidman, W. Bialek, and M. J. Berry. Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 23(37):11539–11553, 2003.

[31] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

[32] T. F. Varley. Decomposing past and future: Integrated information decomposition based on shared probability mass exclusions. *Plos one*, 18(3):e0282950, 2023.

[33] T. F. Varley. Generalized decomposition of multivariate information. *arXiv preprint arXiv:2309.08003*, 2023.

[34] T. F. Varley, M. Pope, M. Grazia, Joshua, and O. Sporns. Partial entropy decomposition reveals higher-order information structures in human brain activity. *Proceedings of the National Academy of Sciences*, 120(30):e2300888120, 2023.

[35] S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

[36] P. L. Williams. *Information dynamics: Its theory and application to embodied cognitive systems*. PhD thesis, Indiana University, 2011.

[37] P. L. Williams and R. D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Appendix A

Proof of work

A.1 Proof of the completeness of Definition 1.

In this part, we will show that do-operation's output is a probability distribution with the same marginal distribution as its input.

Lemma 13. *For $\mathcal{D}_{X,Y,Z}$ and $\mathcal{D}_C$ as in Definition 1, the output $\Pr(A, B, C = x, y, z)$ of (3.1) describes a probability distribution, i.e.,*

$$0 \leq \Pr(A, B, C = x, y, z) \leq 1, \text{ and } \sum_{x,y,z \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}} \Pr(A, B, C = x, y, z) = 1.$$

Furthermore, the marginal distribution $\mathcal{D}_C$ of the output $\mathcal{D}_{A,B,C}$ is equal to the input (call it $\mathcal{D}_{C'}$), i.e.,.

$$\sum_{xy \in \mathcal{X}\mathcal{Y}} \Pr(A, B, C = x, y, z) = \Pr(C' = z).$$

Proof. We begin by showing $0 \leq \Pr(A, B, C = (x, y, z)) \leq 1$. By Definition 1,

$$\Pr(A, B, C = (x, y, z)) = \Pr((X, Y, Z) = (x, y, z)) \cdot \Pr(C = z) / \Pr(Z = z) \tag{A.1}$$

Since $\Pr(Z = z) > 0$ if $\Pr((X, Y, Z) = (x, y, z)) > 0$, (A.1) is well -defined (no zero division). Then, we can write the quotient of the joint probability $\Pr((X, Y, Z) = (x, y, z))$ and the marginal probability $\Pr(Z = z)$ as the conditional probability $\Pr((X, Y) = (x, y) | Z = z)$

and thus

$$(\text{A.1}) = \Pr((X,Y) = (x,y)|Z = z) \cdot \Pr(C = z). \tag{A.2}$$

Since both terms in (A.2) are between 0 and 1, so is $\Pr(A,B,C = (x,y,z))$.

We continue by showing that $\sum_{x,y,z} \Pr((A,B,C) = (x,y,z)) = 1$. Since $\Pr(A,B,C = (x,y,z))$ can be written as (A.2), it follows that

$$\begin{aligned}
&\sum_{x,y,z} \Pr((A,B,C) = (x,y,z)) = \sum_{x,y,z} \Pr((X,Y) = (x,y)|Z = z) \cdot \Pr(C = z) \\
&= \sum_{z \in \mathcal{Z}} \Pr(C = z) \sum_{x,y \in \mathcal{X},\mathcal{Y}} \Pr((X,Y) = (x,y)|Z = z) = \sum_{z \in \mathcal{Z}} \Pr(C = z) = 1.
\end{aligned}$$

Then, we prove that the input $\mathcal{D}_{C'}$ is equal to the marginal distribution $\mathcal{D}_C$ of the output $\mathcal{D}_{A,B,C}$. Since $\mathcal{D}_C$ is the marginal distribution of the output $\mathcal{D}_{A,B,C}$, we have that

$$\Pr(C = z) = \sum_{x,y \in \mathcal{X} \times \mathcal{Y}} \Pr((A,B,C) = (x,y,z)), \tag{A.3}$$

and by Definition 1,

$$\begin{aligned}
(\text{A.3}) &= \sum_{x,y} \Pr((X,Y,Z) = (x,y,z)) \cdot \frac{\Pr(C' = z)}{\Pr(Z = z)} \\
&= \sum_{x,y} \Pr((X,Y) = (x,y)|Z = z) \cdot \Pr(C' = z) \\
&= \Pr(C' = z).
\end{aligned}$$

□

A.2 Proof of Lemma 1.

Proof. Consider the r.h.s of Lemma 1,

$$\sum_{y \in \mathcal{Y}} \Pr(Y = y) H(A_y|C_y). \tag{A.4}$$

By the definition of conditional entropy, we have

$$(\text{A.4}) = \sum_{y\in\mathcal{Y}} \Pr(Y=y) \cdot \sum_{z\in\mathcal{Z}} \Pr(C_y = z) H(A_y | C_y = z) \tag{A.5}$$

By recalling that $\Pr(C_y = z) = \Pr(Z = z|Y = y)$, and by the definition of conditional entropy, we have

$$\begin{aligned}
(\text{A.5}) &= \sum_{y\in\mathcal{Y}} \Pr(Y=y) \sum_{z\in\mathcal{Z}} \Pr(Z=z|Y=y) \cdot (-\sum_{x\in\mathcal{X}} \Pr(A_y = x|C_y = z) \log \Pr(A_y = x|C_y = z)) \\
&= \sum_{y\in\mathcal{Y}} \sum_{z\in\mathcal{Z}} \Pr(Z=z, Y=y) \cdot (-\sum_{x\in\mathcal{X}} \Pr(A_y = x|C_y = z) \log \Pr(A_y = x|C_y = z)) \\
&= \sum_{y\in\mathcal{Y}} \sum_{z\in\mathcal{Z}} \Pr(Z=z, Y=y) \cdot (-\sum_{x\in\mathcal{X}} \frac{\Pr(A_y = x, C_y = z)}{\Pr(C_y = z)} \log \frac{\Pr(A_y = x, C_y = z)}{\Pr(C_y = z)})
\end{aligned} \tag{A.6}$$

Now, notice that summation of (3.1) over all $y \in \mathcal{Y}$ results in

$$\Pr(A_y = x, C_y = z) = \frac{\Pr(X = x, Z = z) \Pr(Z = z|Y = y)}{\Pr(Z = z)},$$

and therefore we have

$$\begin{aligned}
(\text{A.6}) = &\sum_{y\in\mathcal{Y}} \sum_{z\in\mathcal{Z}} \Pr(Z=z, Y=y) \\
&\cdot (-\sum_{x\in\mathcal{X}} \frac{\Pr(X = x, Z = z) \Pr(Z = z|Y = y)}{\Pr(Z = z) \Pr(Z = z|Y = y)} \log \frac{\Pr(X = x, Z = z) \Pr(Z = z|Y = y)}{\Pr(Z = z) \Pr(Z = z|Y = y)})
\end{aligned} \tag{A.7}$$

and hence,

$$\begin{aligned}
(\text{A.7}) &= \sum_{y\in\mathcal{Y}} \sum_{z\in\mathcal{Z}} \Pr(Z=z, Y=y) \cdot (-\sum_{x\in\mathcal{X}} \Pr(X = x|Z = z) \log \Pr(X = x|Z = z)) \\
&= \sum_{z\in\mathcal{Z}} \Pr(Z = z) H(X|Z = z) = H(X|Z).
\end{aligned}$$

□

A.3 Proof of the completeness of Definition 5.

In this part, we will show that $A_{Z|Y}$ is a probability distribution.

Lemma 14. *For $\mathcal{D}_{X,Y,Z}$ in Def. 5, the output $\Pr(A_{Z|Y} = x)$ of (3.4) describes a probability distribution, i.e.,*

$$0 \leq \Pr(A_{Z|Y} = x) \leq 1, \text{ and } \sum_{x \in \mathcal{X}} \Pr(A_{Z|Y} = x) = 1.$$

Proof. Recall that

$$\Pr(A_{Z|Y} = x) = \sum_{y \in \mathcal{Y}} \Pr(Y = y) \Pr(A_y = x). \tag{A.8}$$

By viewing each A_y as the marginal of (A_y, B_y, C_y), we have that

$$(\text{A.8}) = \sum_{y \in \mathcal{Y}} \Pr(Y = y) \sum_{y', z} \Pr(A_y = x, B_y = y', C_y = z), \tag{A.9}$$

which by Definition 1 implies

$$(\text{A.9}) = \sum_{y \in \mathcal{Y}} \Pr(Y = y) \sum_{y', z} (\Pr(X = x, Y = y', Z = z) \cdot \Pr(Z = z | Y = y) / \Pr(Z = z)), \tag{A.10}$$

and by summing over all y' we have

$$\begin{aligned}
(\text{A.10}) &= \sum_{y\in\mathcal{Y}} \Pr(Y=y) \sum_{z\in\mathcal{Z}} (\Pr(X=x, Z=z) \cdot \Pr(Z=z|Y=y) / \Pr(Z=z)) \\
&= \sum_{y,z\in\mathcal{Y}\times\mathcal{Z}} \frac{\Pr(X=x, Z=z) \cdot \Pr(Z=z, Y=y)}{\Pr(Z=z)} \\
&= \sum_{y,z\in\mathcal{Y}\times\mathcal{Z}} \Pr(X=x, Z=z) \cdot \Pr(Y=y|Z=z) \\
&= \sum_{z\in\mathcal{Z}} \Pr(X=x, Z=z) \sum_{y\in\mathcal{Y}} \Pr(Y=y|Z=z) \\
&= \sum_{z\in\mathcal{Z}} \Pr(X=x, Z=z) \\
&= \Pr(X=x).
\end{aligned}$$

Therefore, we have shown $\Pr(A_{Z|Y}=x)$ is a probability distribution and $H(A_{Z|Y}) = H(X)$. □

A.4 Proof of Lemma 3.

Proof. In Corollary 2 we have that $H(A_y) = H(A_{Z|Y}|Y=y)$. Therefore, we can write the left hand side of Equation (3.5) as

$$\sum_{y\in\mathcal{Y}} \Pr(Y=y) H(A_{Z|Y}|Y=y).$$

By Lemma 2, we have $H(A_{Z|Y}) = H(X)$. Then by the definition of conditional entropy, we have:

$$\sum_{y\in\mathcal{Y}} \Pr(Y=y) H(A_{Z|Y}|Y=y) = H(A_{Z|Y}|Y) \leq H(A_{Z|Y}) = H(X). \qquad \square$$

A.5 Proof of Lemma 4.

Proof. By Corollary 1, Lemma 4 is equivalent to:

$$\sum_{y\in\mathcal{Y}} \Pr(Y=y)H(A_y) - H(X|Z) \le I(X,Z), \text{ which is } \sum_{y\in\mathcal{Y}} \Pr(Y=y)H(A_y) \le H(X). \tag{A.11}$$

Since (A.11) coincides with the statement of Lemma 3, the proof of Lemma 4 follows. □

A.6 Proof of Lemma 5.

Proof. By Definition 2,

$$\mathrm{Un}(X \to Z|Y) = \sum_{y\in\mathcal{Y}} \Pr(Y=y) I(A_y; C_y). \tag{A.12}$$

Since the mutual information $I(A_yC_y)$ is less than the entropy $H(C_y)$, we have:

$$\mathrm{Un}(X \to Z|Y) \le \sum_{y\in\mathcal{Y}} \Pr(Y=y) H(C_y) \tag{A.13}$$

By the definition of entropy, we have (A.13) equals:

$$\sum_{y\in\mathcal{Y}} \Pr(Y=y) \sum_{z\in\mathcal{Z}} \Pr(C_y=z)(-\log\Pr(C_y=z)) \tag{A.14}$$

By Definition 2 that $\Pr(C_y=z) = \Pr(Z=z|Y=y)$, we have (A.14) equals:

$$\sum_{y\in\mathcal{Y}} \Pr(Y=y) \sum_{z\in\mathcal{Z}} \Pr(Z=z|Y=y)(-\log\Pr(Z=z|Y=y)) \tag{A.15}$$

which is $H(Z|Y)$ by the definition of conditional entropy. □

A.7 Proof of Lemma 6.

Proof. By Definition 3, we have:

$$\mathrm{Red}(X, Y \to Z) = I(X, Z) - \mathrm{Un}(X \to Z|Y) \tag{A.16}$$

By adding and then subtracting the conditional entropy $H(X|Z)$, we obtain (A.16) equals:

$$(I(X, Z) + H(X|Z)) - (\mathrm{Un}(X \to Z|Y) + H(X|Z)) \tag{A.17}$$

Since mutual information plus conditional entropy equals information entropy, (A.17) can be written as:

$$H(X) - (\mathrm{Un}(X \to Z|Y) + H(X|Z)) \tag{A.18}$$

Since Corollary 1 states:

$$\mathrm{Un}(Y \to Z|X) = \sum_{y \in \mathcal{Y}} \Pr(Y = y) H(A_y) - H(X|Z), \tag{A.19}$$

it follows that

$$(\text{A.18}) = H(X) - \sum_{y \in \mathcal{Y}} \Pr(Y = y) H(A_y). \tag{A.20}$$

Recall that $H(A_{Z|Y}|Y = y) = H(A_y)$ for all $y \in \mathcal{Y}$ by Corollary 2, and that $H(X) = H(A_{Z|Y})$ by Lemma 2. Therefore,

$$(\text{A.20}) = H(A_{Z|Y}) - \sum_{y \in \mathcal{Y}} \Pr(Y = y) H(A_{Z|Y}|Y = y) = H(A_{Z|Y}) - H(A_{Z|Y}|Y) = I(A_{Z|Y}; Y). \square$$

A.8 Proof of Lemma 7.

Proof. To prove that $I(A_{Z|Y}; Y) = I(B_{Z|X}; X)$, it suffices to show that $\Pr(A_{Z|Y} = x, Y = y) = \Pr(X = x, B_{Z|X} = y)$ for all $x, y \in \mathcal{X} \times \mathcal{Y}$.

Indeed, for every $x, y \in \mathcal{X} \times \mathcal{Y}$ we have that

$$\Pr(A_{Z|Y} = x, Y = y) = \Pr(Y = y) \Pr(A_{Z|Y} = x | Y = y) = \Pr(Y = y) \Pr(A_y = x), \quad \text{(A.21)}$$

where the last transition follows from Definition 5. Then, by considering A_y as a marginal of (A_y, B_y, C_y), we have that

$$\text{(A.21)} = \Pr(Y = y) \sum_{y', z} \Pr(A_y = x, B_y = y', C_y = z) \quad \text{(A.22)}$$

and then by Definition 1,

$$\text{(A.22)} = \Pr(Y = y) \sum_{y', z} (\Pr(X = x, Y = y', Z = z) \cdot \Pr(Z = z | Y = y) / \Pr(Z = z)). \quad \text{(A.23)}$$

By summing over all y' we have

$$\begin{aligned} \text{(A.23)} &= \Pr(Y = y) \sum_{z \in \mathcal{Z}} (\Pr(X = x, Z = z) \cdot \Pr(Z = z | Y = y) / \Pr(Z = z)) \\ &= \sum_{z \in \mathcal{Z}} \Pr(X = x, Z = z) \cdot \frac{\Pr(Y = y, Z = z)}{\Pr(Z = z)} \end{aligned} \quad \text{(A.24)}$$

Now, the proof will be concluded by following similar steps to (A.21)-(A.24), only in reversed order.

$$\begin{aligned} \text{(A.24)} &= \sum_{z \in \mathcal{Z}} \frac{\Pr(Z = z, X = x)}{\Pr(Z = z)} \Pr(Y = y, Z = z) = \Pr(X = x) \\ &\quad \cdot \sum_{z \in \mathcal{Z}} (\Pr(Y = y, Z = z) \cdot \Pr(Z = z | X = x) / \Pr(Z = z)) \\ &= \Pr(X = x) \sum_{x', z \in \mathcal{X} \times \mathcal{Z}} (\Pr(X = x', Y = y, Z = z) \cdot \Pr(Z = z | X = x) / \Pr(Z = z)) \\ &= \Pr(X = x) \sum_{x', z \in \mathcal{X}, \mathcal{Z}} \Pr(A_x = x', B_x = y, C_x = z) \\ &= \Pr(X = x) \Pr(B_x = y) = \Pr(X = x) \Pr(B_{Z|X} = y | X = x) \\ &= \Pr(X = x, B_{Z|X} = y). \end{aligned}$$

□

A.9 Proof of Lemma 8.

Proof. By Definition 2, we have:

$$\mathrm{Un}((X,X')\to(Z,Z')|(Y,Y')) = \sum_{y,y'\in\mathcal{Y}\times\mathcal{Y}} (\Pr(Y,Y'=y,y')I(A_y,A'_{y'};C_y,C'_{y'})) \tag{A.25}$$

To show the additivity property of (A.25), we begin by showing that A_y, B_y, C_y are independent of $A'_{y'}, B'_{y'}, C'_{y'}$ for every $y, y' \in \mathcal{Y}\times\mathcal{Y}$, where

$$\mathcal{D}_{(A_y,B_y,C_y)} = do(\mathcal{D}_{X,Y,Z}|\mathcal{D}_{C_y}),\ \text{and}\mathcal{D}_{(A'_{y'},B'_{y'},C'_{y'})} = do(\mathcal{D}_{X',Y',Z'}|\mathcal{D}_{C'_{y'}}).$$

To this end, by Definition 1 and Definition 2, for every $y, y', \hat{y}, \hat{\hat{y}} \in \mathcal{Y}$, every $x, x' \in \mathcal{X}$, and every $z, z' \in \mathcal{Z}$, we have:

$$\begin{aligned}
&\Pr((A_y,A'_{y'}),(B_y,B'_{y'}),(C_y,C'_{y'}) = (x,x'),(\hat{y},\hat{\hat{y}}),(z,z')) \\
=&(\Pr((X,X'),(Y,Y'),(Z,Z') = (x,x'),(\hat{y},\hat{\hat{y}}),(z,z')) \\
&\cdot \Pr((Z=z,Z'=z')|(Y=y,Y'=y'))/\Pr(Z=z,Z'=z'))
\end{aligned} \tag{A.26}$$

Since X, Y, Z are independent of X', Y', Z', for every $x, x \in \mathcal{X}$ every $y, y' \in \mathcal{Y}$, and every $z, z' \in \mathcal{Z}$, we have:

$$\begin{aligned}
&\Pr((X,X'),(Y,Y'),(Z,Z') = (x,x'),(\hat{y},\hat{\hat{y}}),(z,z')) \\
&= \Pr(X,Y,Z=x,\hat{y},z)\cdot\Pr(X',Y',Z'=x',\hat{\hat{y}},z'),
\end{aligned}$$

$$\begin{aligned}
\Pr(Z=z,Z'=z'|Y=y,Y'=y') &= \Pr(Z=z|Y=y,Y'=y')\cdot\Pr(Z'=z'|Y=y,Y'=y') \\
&= \Pr(Z=z|Y=y)\cdot\Pr(Z'=z'|Y'=y'),
\end{aligned}$$

$$\text{and}\ \ \Pr(Z=z,Z'=z') = \Pr(Z=z)\cdot\Pr(Z'=z').$$

Therefore,

$$\begin{aligned}(\text{A.26}) &= \Pr(X,Y,Z=x,\hat{y},z)\cdot\Pr(Z=z|Y=y)/\Pr(Z=z)\\ &\quad\cdot\Pr(X',Y',Z'=x',\hat{\hat{y}},z')\cdot\Pr(Z'=z'|Y'=y')/\Pr(Z'=z')\\ &= \Pr(A_y=x,B_y=\hat{y},C_y=z)\cdot\Pr(A'_{y'}=x',B'_{y'}=\hat{\hat{y}},C'_{y'}=z')\end{aligned}$$

Therefore, we have that A_y, B_y, C_y are independent of $A'_{y'}, B'_{y'}, C'_{y'}$. Coming back to (A.25), we have:

$$\begin{aligned}(\text{A.25}) &= \sum_{y,y'\in\mathcal{Y}\times\mathcal{Y}}\Pr(Y=y)\cdot\Pr(Y'=y')\cdot(I(A_y;C_y)+I(A'_{y'};C'_{y'}))\\ &= \sum_{y,y'\in\mathcal{Y}\times\mathcal{Y}}\Pr(Y=y)\Pr(Y'=y')I(A_y;C_y)+\sum_{y,y'\in\mathcal{Y}\times\mathcal{Y}}\Pr(Y'=y')\Pr(Y=y)I(A'_{y'};C'_{y'})\\ &= \sum_{y\in\mathcal{Y}}\Pr(Y=y)I(A_y;C_y)+\sum_{y'\in\mathcal{Y}}\Pr(Y'=y')I(A'_{y'};C'_{y'})\\ &= \mathrm{Un}(X\to Z|Y)+\mathrm{Un}(X'\to Z'|Y').\end{aligned}$$

□

A.10 Proof of Lemma 9.

Proof. Recall that Lemma 6 states that $\mathrm{Red}(X,Y\to Z)=I(A_{Z|Y};Y)$. Therefore, since $I(A_{Z|Y};Y)$ is a continuous function of $\mathcal{D}_{A_{Z|Y},Y}$, it suffices to prove that the mapping $\mathcal{F}(\mathcal{D}_{X,Y,Z})=\mathcal{D}_{A_{Z|Y},Y}$, that is implied by Definition 1, Definition 2, and Definition 5, is continuous.

Specifically, let $\mathcal{F}:\Delta_1\to\Delta_2$ be the implicit mapping mentioned above, where Δ_1 (resp. Δ_2) is the suitable probability simplex, i.e., the set of all tensors in $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|\times|\mathcal{Z}|}$ (resp. $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$) with nonnegative entries which sum to 1.

To show that $\mathcal{F}$ is continuous, we show that

$$\lim_{w\to w_0}\mathcal{F}(w)=\mathcal{F}(w_0)$$

for every $w_0\in\Delta_1$. Indeed, by Definition 5, $\mathcal{D}_{A_{Z|Y},Y}=\mathcal{F}(\mathcal{D}_{X,Y,Z})$ satisfies

$$\Pr(A_{Z|Y},Y=x,y)=\Pr(Y=y)\Pr(A_y=x)\tag{A.27}$$

for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, where A_y is the marginal of $\mathcal{D}_{A_y,B_y,C_y} = do(\mathcal{D}_{X,Y,Z}|\mathcal{D}_{Z|Y=y})$ given in Definition 2. Therefore,

$$\begin{aligned}(\text{A.27}) &= \Pr(Y=y) \sum_{y',z \in \mathcal{Y}\times\mathcal{Z}} \Pr(A_y, B_y, C_y = x, y', z) \\ &= \sum_{y',z \in \mathcal{Y}\times\mathcal{Z}} \Pr(Y=y) \Pr(A_y, B_y, C_y = x, y', z), \end{aligned} \tag{A.28}$$

for every $x \in \mathcal{X}$ and every $y \in \mathcal{Y}$.

Since continuity is preserved by summation, to show the continuity of (A.28), it suffices to show the continuity of each term in (A.28). To this end, for $x \in \mathcal{X}$, $y, y' \in \mathcal{Y}$, and $z \in \mathcal{Z}$, let $\mathcal{G}_{x,y,y',z}$ be the mapping from Δ_1 to $\mathbb{R}$ which is implied by the respective term in (A.28), i.e.,

$$\mathcal{G}_{x,y,y',z}(w) = \Pr{}_w(Y=y) \Pr{}_w(A_y, B_y, C_y = x, y', z) \tag{A.29}$$

for every $w \in \Delta_1$, where $\Pr_w(E)$ is the probability of event E implied by the joint distribution on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ given by the tensor w. Then, it suffices to show that

$$\lim_{w \to w_0} \mathcal{G}_{x,y,y',z}(w) = \mathcal{G}_{x,y,y',z}(w_0)$$

for every $w_0 \in \Delta_1$.

By Definition 1, we have:

$$(\text{A.29}) = \begin{cases} 0, \text{ if } \Pr_w(Z=z) = 0 \text{ or} \\ \Pr_w(Y=y) \Pr_w(X, Y, Z = x, y', z) \cdot \quad \Pr_w(Z=z|Y=y) / \Pr_w(Z=z), \text{ otherwise,} \end{cases} \tag{A.30}$$

$$= \begin{cases} 0, \text{ if } \Pr_w(Z=z) = 0, \text{or} \\ \Pr_w(X, Y=x, y'|Z=z) \Pr_w(Y, Z=y, z) \end{cases} \tag{A.31}$$

for every $x \in \mathcal{X}$, every $y, y' \in \mathcal{Y}$, and every $z \in \mathcal{Z}$. Next, we split to cases.

Case 1: $\Pr_{w_0}(Z=z) > 0$. There exists a neighborhood around w_0 for which $\Pr_w(Z=z) > 0$ for every w in that neighborhood. Here, continuity can be explained in simple terms as a

composition of continuous functions, i.e., the product of a joint probability $\Pr_w(Y, Z = y, z)$ and a conditional joint probability $\Pr_w(X, Y = x, y' | Z = z)$.

Case 2: $\Pr_{w_0}(Z = z) = 0$. Observe that for this w_0 we have $\mathcal{G}_{x,y,y',z}(w_0) = 0$ by (A.30), and thus we wish to show that

$$\lim_{w \to w_0} \mathcal{G}_{x,y,y',z}(w) = \mathcal{G}_{x,y,y',z}(w_0) = 0 \tag{A.32}$$

for every $w_0 \in \Delta_1$. First, we can without loss of generality assume that $\Pr_w(Z = z) > 0$ for every w in a neighborhood of w_0, since all w for which $\Pr_w(Z = z) = 0$ already have that $\mathcal{G}_{x,y,y',z}(w) = 0$.

It remains to show that $\lim_{w \to w_0} \mathcal{G}_{x,y,y',z}(w) = 0$ under the condition that $\Pr_w(Z = z) > 0$ for every w in the limit operation. To show that, write

$$\begin{aligned} \lim_{w \to w_0} \mathcal{G}_{x,y,y',z}(w) &\overset{(A.30)}{=} \lim_{w \to w_0} \Pr_w(X, Y = x, y' | Z = z) \cdot \Pr_w(Y, Z = y, z) \\ &= \lim_{w \to w_0} \Pr_w(X, Y = x, y' | Z = z) \cdot \textstyle\sum_{\hat{x}} w(\hat{x}, y, z). \end{aligned} \tag{A.33}$$

Now, observe that $\Pr_{w_0}(Z = z) = 0$ implies that $w_0(\hat{x}, \hat{y}, z) = 0$ for every $\hat{x} \in \mathcal{X}$ and every $\hat{y} \in \mathcal{Y}$. Therefore, it follows that $\lim_{w \to w_0} w(\hat{x}, y, z) = w_0(\hat{x}, y, z) = 0$ for every $\hat{x} \in \mathcal{X}$. Hence,

$$\lim_{w \to w_0} \sum_{\hat{x} \in \mathcal{X}} w(\hat{x}, y, z) = \sum_{\hat{x} \in \mathcal{X}} \lim_{w \to w_0} w(\hat{x}, y, z) = 0.$$

Therefore, (A.33) is a limit of a bounded quantity and a quantity which goes to zero, and hence equals zero itself, which concludes the proof. □

A.11 Proof of Lemma 10.

Proof. Since $Z = (X, Y)$, we identify the alphabet $\mathcal{Z}$ of Z as $\mathcal{X} \times \mathcal{Y}$, and as a result, observe that

$$\Pr(X, Y, Z = x, y, (x', y')) = 0 \text{ whenever } x \neq x' \text{ or } y \neq y',$$

for every $x, x' \in \mathcal{X}$ and every $y, y' \in \mathcal{Y}$.

Similarly, it follows from Definition 1 and Definition 2 that $\Pr(X, Y, Z = x, y, z) = 0$ implies that $\Pr(A, B, C = x, y, z) = 0$ for every $x, y, z \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, and hence

$$\Pr(A_y, B_y, C_y = x, y', (x', y'')) = 0 \tag{A.34}$$

whenever $x \neq x'$ or $y' \neq y''$, for every $x, x' \in \mathcal{X}$, and every $y, y', y'' \in \mathcal{Y}$.

Also, since the marginal distribution $\mathcal{D}_{C_y}$ of $\mathcal{D}_{A_y,B_y,C_y} = do(\mathcal{D}_{X,Y,Z}|\mathcal{D}_{Z|Y=y})$ is identical to $\mathcal{D}_{Z|Y=y}$ (by the "furthermore" part of Lemma 13) it follows that

$$\Pr(C_y = (x, y')) = \Pr(Z = (x, y')|Y = y) \tag{A.35}$$

for every $y, y' \in \mathcal{Y}$. Furthermore, considering that $Z = (X, Y)$, (A.35) also implies that

$$\Pr(C_y = (x, y')) = 0 \text{ whenever } y \neq y', \tag{A.36}$$

for every $x \in \mathcal{X}$ and every $y, y' \in \mathcal{Y}$. When $y' = y$, however, for every $x \in \mathcal{X}$ and every $y \in \mathcal{Y}$ we have

$$\Pr(C_y = (x, y)) = \Pr(Z = (x, y)|Y = y) = \Pr(X = x|Y = y) = \Pr(X = x), \tag{A.37}$$

where the latter two steps follow since $Z = (X, Y)$ and $I(X, Y) = 0$. Besides, (A.36) and (A.34) imply that

$$\Pr(A_y, B_y, C_y = x, y', (x', y'')) = 0 \text{ whenever } x' \neq x, \text{ or } y' \neq y, \text{ or } y'' \neq y, \tag{A.38}$$

for every $x, x' \in \mathcal{X}$ and every $y, y', y'' \in \mathcal{Y}$.

We now turn to prove that

$$\Pr(A_y, C_y = x, (x, y)) = \Pr(A_y = x) = \Pr(C_y = (x, y))$$

for every $x \in \mathcal{X}$ and every $y \in \mathcal{Y}$, by showing that each of the three expressions equals an identical common expression, which in turn is shown to be equal $H(X)$.

We begin with $\Pr(A_y, C_y = x, (x, y))$, for which every $x, x' \in \mathcal{X}$ and every $y, y' \in \mathcal{Y}$ satisfy

$$\Pr(A_y, C_y = x, (x', y')) = \sum_{y'' \in \mathcal{Y}} \Pr(A_y, B_y, C_y = x, y'', (x', y')). \tag{A.39}$$

By (A.38), each summand in (A.39) equals zero if either $x' \neq x$ or $y' \neq y$ or $y'' \neq y$, and hence

$$(A.39) = \begin{cases} 0, \text{ if } x' \neq x \text{ or } y' \neq y \\ \Pr(A_y, B_y, C_y = x, y, (x, y)), \text{ otherwise.} \end{cases} \tag{A.40}$$

Similarly, for $\Pr(A_y = x)$, every $x \in \mathcal{X}$ and every $y \in \mathcal{Y}$ satisfy

$$\Pr(A_y = x) = \sum_{y'', (x', y')} \Pr(A_y, B_y, C_y = x, y'', (x', y')) \overset{(A.38)}{=} \Pr(A_y, B_y, C_y = x, y, (x, y)). \tag{A.41}$$

Further, for $\Pr(C_y = (x, y'))$, every $x \in \mathcal{X}$ and every $y, y' \in \mathcal{Y}$ satisfy

$$\Pr(C_y = (x, y')) = \sum_{x', y''} \Pr(A_y, B_y, C_y = x', y'', (x, y')) \overset{(A.38)}{=} \Pr(A_y, B_y, C_y = x, y, (x, y')), \tag{A.42}$$

and observe that according to (A.38), Eq. (A.42) is equal to zero whenever $y \neq y'$.

Therefore, by (A.40), (A.41), and (A.42), we have

$$\Pr(A_y, B_y, C_y = x, y'', (x', y')) = \begin{cases} 0, \text{ if } x' \neq x \text{ or } y' \neq y \text{ or } y'' \neq y, \text{and otherwise} \\ \Pr(A_y, C_y = x, (x, y)) = \Pr(A_y = x) = \Pr(C_y = (x, y)). \end{cases} \tag{A.43}$$

for every $x, x' \in \mathcal{X}$ and for every $y, y', y'' \in \mathcal{Y}$.

Finally, we show that $H(A_y, C_y) = H(A_y) = H(C_y) = H(X)$. By the definition of entropy,

$$H(A_y, C_y) = \sum_{x,(x',y')} \Pr(A_y, C_y = x, (x', y') \cdot \log(1/\Pr(A_y, C_y = x, (x', y'))) \tag{A.44}$$

Since by (A.40), every term in the summation in (A.44) with $x' \neq x$ or $y' \neq y$ equals zero, and otherwise $\Pr(A_y, C_y = x, (x, y)) = \Pr(A_y, B_y, C_y = x, y, (x, y))$, we have that

$$(A.44) = \sum_x \Pr(A_y, B_y, C_y = x, y, (x, y)) \log(1/\Pr(A_y, B_y, C_y = x, y, (x, y))). \tag{A.45}$$

Similarly, we have

$$\begin{aligned} H(A_y) &= \sum_x \Pr(A_y = x) \log(1/\Pr(A_y = x)) \\ &\overset{(A.41)}{=} \sum_x (\Pr(A_y, B_y, C_y = x, y, (x, y)) \log(1/\Pr(A_y, B_y, C_y = x, y, (x, y))), \end{aligned} \tag{A.46}$$

and

$$\begin{aligned} H(C_y) &= \sum_{(x,y')} \Pr(C_y = (x, y')) \log(1/\Pr(C_y = (x, y'))) \\ &\overset{(A.42)}{=} \sum_x \Pr(A_y, B_y, C_y = x, y, (x, y)) \log(1/\Pr(A_y, B_y, C_y = x, y, (x, y))). \end{aligned} \tag{A.47}$$

Further, for every $y \in \mathcal{Y}$, we have

$$\begin{aligned} H(X) &= \sum_x \Pr(X = x) \log(1/\Pr(X = x)) \overset{(A.37)}{=} \sum_x \Pr(C_y = (x, y)) \log(1/\Pr(C_y = (x, y))) \\ &= \sum_x \Pr(A_y, B_y, C_y = x, y, (x, y))(1/\Pr(A_y, B_y, C_y = x, y, (x, y))). \end{aligned} \tag{A.48}$$

Then, since (A.45), (A.46), (A.47), and (A.48) are all equal, it follows that $H(A_y, C_y) = H(A_y) = H(C_y) = H(X)$ for all $y' \in \mathcal{Y}$.

To conclude the proof, recall Definition 3 of Red,

$$\mathrm{Red}(X, Y \to Z) = I(X; Z) - \mathrm{Un}(X \to Z|Y) \overset{(2)}{=} I(X; Z) - \sum_{y \in \mathcal{Y}} \Pr(Y = y) I(A_y; C_y). \tag{A.49}$$

Also, observe that $I(A_y; C_y) = H(A_y) + H(C_y) - H(A_y, C_y) = 2H(X) - H(X) = H(X)$ for all $y \in \mathcal{Y}$ according to the above discussion, and that $I(X; Z) = H(X)$ since $I(X; Y) = 0$ and $Z = (X, Y)$. Then, by (A.49), we have $\text{Red}(X, Y \to Z) = 0$, which completes the proof. □

A.12 Proof of Theorem 1

Proof. Suppose we have a multivariate system containing a target variable Y and source variables $X_1, \cdots, X_n$. For the convenience of expression, we use $\mathcal{X}$ to represent all the source variables $X_1, \cdots, X_n$. The proof is to show that $Red(Y : \mathcal{X}, Y) = Red(Y; \mathcal{X})$ and $Red(U : \mathcal{X}, Y) = Red(Y : \mathcal{X}, Y)$, where U is the union variable of Y and $\mathcal{X}$, such that $U = (\mathcal{X}, Y)$. (The entropy of the union variable U can be expressed as $H(U) = H(\mathcal{X}, Y)$.) Then, we can demonstrate that redundant information is equal regardless of which variable is chosen as the target variable.

Step One, to prove $Red(Y : \mathcal{X}, Y) = Red(Y : \mathcal{X})$:

By Definition 6,

$$Red(Y : \mathcal{X}) = \sup_Q \{I(Q : Y) : Q \sqsubset_Y X_i, \forall i \in \{1 \cdots n\}\} \tag{A.50}$$

According to Blackwell order, $Q \sqsubset_Y Y$, since Y has all of the information about Y. Then, we have:

$$\sup_Q \{I(Q : Y) : Q \sqsubset_Y X_i, \forall i \in \{1 \cdots n\}\} = \sup_Q \{I(Q : Y) : Q \sqsubset_Y Y, Q \sqsubset_Y X_i, \forall i \in \{1 \cdots n\}\} \tag{A.51}$$

Therefore, $Red(Y : \mathcal{X}, Y) = Red(Y; \mathcal{X})$.

Step Two, to prove $Red(U : \mathcal{X}, Y) = Red(Y : \mathcal{X}, Y)$:

Building upon the conclusion that $Red(Y : \mathcal{X}, Y) = Red(Y : \mathcal{X})$, we can replace the target variable with the union variable $U = (\mathcal{X}, Y)$.

By Definition 6,

$$Red(U:\mathcal{X},Y)=\sup_{Q}\{I(Q:U):Q\sqsubset_U Y,Q\sqsubset_U X_i,\forall i\in\{1\cdots n\}\} \tag{A.52}$$

Let Q^* satisfies or infinitely approaches the above conditions:

$$\begin{aligned}I(Q^*:U)&=Red(U:\mathcal{X},Y)-\varepsilon,\forall\varepsilon>0\\&=\sup_{Q}\{I(Q:U):Q\sqsubset_U Y,Q\sqsubset_U X_i,\forall i\in\{1\cdots n\}\}-\varepsilon,\forall\varepsilon>0,\end{aligned}$$

Since $U=(\mathcal{X},Y)(H(Y|U)=0)$, then $I(Q^*:U)\geq I(Q^*:Y)$. Considering that $Q^*\sqsubset_U Y$, which means Y has all of the information that Q^* has, about the target variable U, such that $I(Q^*:U)\leq I(Q^*:Y)$, we have:

$$I(Q^*:U)=I(Q^*:Y) \tag{A.53}$$

Since Y has all the information about itself, we have:

$$Q^*\sqsubset_Y Y \tag{A.54}$$

Since $U=(\mathcal{X},Y)(H(Y|U)=0)$ and $Q^*\sqsubset_U X_i,\forall i\in\{1\cdots n\}\}$ (X_i has all of the information that Q^* has, about the target variable U), we have:

$$Q^*\sqsubset_Y X_i,\forall i\in\{1\cdots n\} \tag{A.55}$$

Therefore, by Equation A.52-A.55 and Definition 6, we obtain:

$$Red(U:\mathcal{X},Y)=\sup_{Q}\{I(Q:Y):Q\sqsubset_Y Y,Q\sqsubset_Y X_i,\forall i\in\{1\cdots n\}\}=Red(Y:\mathcal{X},Y)$$

In Summary: Since we have established that $Red(Y:\mathcal{X},Y)=Red(Y:\mathcal{X})$, and $Red(U:\mathcal{X},Y)=Red(Y:\mathcal{X},Y)$, we can conclude that for all X_i in $\{\mathcal{X}\}$, $Red(X_i:Y,\{\mathcal{X}\}\setminus X_i)=$

$Red(Y : \{\mathcal{X}\})$. Therefore, Theorem 1 is proved, and we can use $Red(X_1, \cdots, X_n)$ or $Red_{1\cdots n}$ denote the redundant information within the system $\{X_1, \cdots, X_n\}$. □

A.13 Proof of Lemma 11

Proof. According to Axiom 1, unique information is a part of the information provided by the source variable to the target variable, that is, mutual information minus redundant information. In a three-variable system $\{X_1, X_2, X_3\}$, by Axiom 1,

$$Un(X_i : X_j) = I(X_i; X_j) - Red(X_i : X_j, X_k), \forall i \neq j \in \{1, 2, 3\} \tag{A.56}$$

Since $I(X_i : X_j) = I(X_j : X_i)$ according to the symmetry of Shannon's formula [31], and $Red(X_i : X_j, X_k) = Red(X_j : X_i, X_k) = Red(X_i, X_j, X_k)$ according to Theorem 1, we have:

$$Un(X_i : X_j) = I(X_i; X_j) - Red(X_i : X_j, X_k) = I(X_j; X_i) - Red(X_j : X_i, X_k) = Un(X_j : X_i)$$

For general multivariate systems $X_1, \ldots, X_n$, we can prove the symmetry of unique information between any two variables X_i and X_j by combining other variables $X_1, \cdots, X_n \setminus X_i, X_j$ into one variable X_k. Therefore, we proved the theorem, and we can represent this information atom as $Un(X_i, X_j)$, or $Un_{i,j}$. □

A.14 Proof of Lemma 12

Proof. According to Axiom 1, Lemma 11, and the chain rule of Shannon formula, for a three-variable system with X_i, X_j, X_k:

$$\begin{aligned} Syn(X_k : X_i, X_j) &= H(X_k|X_j) - H(X_k|X_i, X_j) - Un(X_i, X_k) \\ &= (H(X_j, X_k) - H(X_j)) - (H(X_i, X_j, X_k) - H(X_i, X_j)) - Un(X_i, X_k) \\ &= H(X_j, X_k) + H(X_i, X_j) - H(X_j) - H(X_i, X_j, X_k) - Un(X_i, X_k) \\ &= (H(X_i, X_j) - H(X_j)) - (H(X_i, X_j, X_k) - H(X_j, X_k)) - Un(X_i, X_k) \\ &= H(X_i|X_j) - H(X_i|X_j, X_k) - Un(X_i, X_k) \\ &= Syn(X_i : X_j, X_k) \end{aligned}$$

Therefore, we demonstrate that X_i and X_k are interchangeable, and since X_i and X_j are interchangeable as source variables, we proved that all variables are interchangeable. For general multivariate systems $X_1, \cdots, X_n$, we can prove the symmetry of synergistic information between any two variables X_i and X_k by combining other variables into one variable X_j. Therefore, we proved Lemma 12 and we can write synergistic information in the form of $Syn(X_1, \cdots, X_n)$ or $Syn_{1\cdots n}$. □