# A New Framework for Decomposing MUltivariate Information 

Conor Finn

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Centre for Complex Systems<br>The University of Sydney

Submitted September 2019
Revised April 2020

To my family.

## Abstract

What are the distinct ways in which a set of predictor variables can provide information about a target variable? When does a variable provide unique information, when do variables share redundant information, and when do variables combine synergistically to provide complementary information? The redundancy lattice from the partial information decomposition of Williams and Beer provided a promising glimpse at the answer to these questions. However, this structure was constructed using a much-criticised measure of redundant information, and despite sustained research, no completely satisfactory replacement measure has been proposed.

This thesis presents a new framework for information decomposition that is based upon the decomposition of pointwise mutual information rather than mutual information. The framework is derived in two separate ways. The first of these derivations is based upon a modified version of the original axiomatic approach taken by Williams and Beer. However, to overcome the difficulty associated with signed pointwise mutual information, the decomposition is applied separately to the unsigned entropic components of pointwise mutual information which are referred to as the specificity and ambiguity. This yields a separate redundancy lattice for each component. Based upon an operational interpretation of redundancy, measures of redundant specificity and redundant ambiguity are defined which enables one to evaluate the partial information atoms separately for each lattice. These separate atoms can then be recombined to yield the sought-after multivariate information decomposition. This framework is applied to canonical examples from the literature and the results and various properties of the decomposition are discussed. In particular, the pointwise decomposition using specificity and ambiguity is shown to satisfy a chain rule over target variables, which provides new insights into the so-called two-bit-copy example.

The second approach begins by considering the distinct ways in which two marginal observers can share their information with the non-observing individual third party. Several novel measures of information content are introduced, namely the union, intersection and unique information contents. Next, the algebraic structure of these new measures of shared marginal information is explored, and it is shown that the structure of shared marginal information is that of a distributive lattice. Furthermore, by using the fundamental theorem of distributive lattices, it is shown that these new measures are isomorphic to a ring of sets. Finally, by combining this structure together with the semi-lattice of joint information, the redundancy lattice form partial information decomposition is found to be embedded within this larger algebraic structure. However, since this structure considers information contents, it is actually equivalent to the specificity lattice from the first derivation of pointwise partial information decomposition. The thesis then closes with a discussion about whether or not one should combine the information contents from the specificity and ambiguity lattices.

## Declaration of Authorship

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any other degree. I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Conor Finn, September 2019.

## Publications

Parts of this thesis have appeared in the following publications and submitted manuscripts during the candidature for this degree.

1. Finn, C. \& Lizier, J. T. Probability mass exclusions and the directed components of mutual information. Entropy 20, 826 (2018).
2. Finn, C. \& Lizier, J. T. Pointwise partial information decomposition using the specificity and ambiguity lattices. Entropy 20, 297 (2018).
3. Finn, C. \& Lizier, J. T. Generalised measures of multivariate information content. Entropy 22, 216 (2020).
4. Finn, C. \& Lizier, J. T. Quantifying information modification in cellular automata using pointwise partial information decomposition in Artificial Life Conference Proceedings (2018), 386-387.
5. Wollstadt, P. et al. The Information Dynamics Toolkit xl: a Python package for the efficient analysis of multivariate information dynamics in networks. Journal of Open Source Software 4, 1081 (2019).
6. Wibral, M., Finn, C., Wollstadt, P., Lizier, J. \& Priesemann, V. Quantifying information modification in developing neural networks via partial information decomposition. Entropy 19, 494 (2017).

Publications 1-3 comprise the original contribution of this thesis. Publications 4-6 are related work that were contributed during my candidature which demonstrate the applicability of this research.

## Acknowledgements

As a complex systems scientist, I am cognisant of that fact that the ideas presented in this thesis are not mine insomuch as they are an emergent phenomena associated with the dynamics of an entire society. With this in mind, I think it is only appropriate that I first acknowledge both the Irish state for providing me with free tertiary education, and EPSRC for funding my masters level studies at the University of Warwick. Specifically regarding this thesis, I am extremely grateful to the University of Sydney and CSIRO's Data 61 for funding my research over the past three and a half years. I would also like to acknowledge the Austrian Research Council, Universities Australia, the German Academic Exchange Service (DAAD), and the Centre for Complex Systems for providing travel and other miscellaneous funding.

On a more personal level, there are a number of individuals to whom I feel particularly indebted. Firstly, I would like to express my deepest gratitude to my primary supervisor. Joe, you have given me the freedom to explore my own ideas whilst ensuring that I do not wander off into the weeds. I will be ever grateful for your enthusiastic response to a random enquiry from far side of the world, for that idea you had whilst brushing your teeth one night, for surviving the onslaught of misspellings, and for our many discussions regarding information, axioms and Venn diagrams (including where not to draw them). As your first full PhD supervisee, I have only one piece of advice-just keep on doing what you are doing!

Next, I wish to thank my secondary supervisors. Mikhail, our group owes you a great deal for the tireless work you put into supporting us all. Personally, I am most grateful for the guidance you have given me on surviving academia. Not many supervisors would quote the Art of War to their students in their first meeting. More should. Rose, your support from afar was always much appreciated. I also want to acknowledge my many awesome collaborators, especially Michael Wibral, Patricia Wollstadt, Viola Priesemann and Pedro Mediano.

This list of acknowledgements would be complete without mentioning friends at the University of Sydney. I would particularly like to thank Emanuele, Nathan, Leonardo, David, Richard and Oliver for our many conversations that alternated haphazardly between genuine science, surreal politics and the disturbingly arcane. I must also thank my many non-academic friends, both here in Sydney and back in Ireland, who remind me that most people have other interests.

Of course, I must thank my family. Mary, your unwavering encouragement has pushed me further than anything else, even when certain teachers did not see it that way. I cannot tell you how unbelievable appreciative I am for this. Ciarán, I owe my initial interest in science and mathematics to you and the many non-fiction books you first read to me and later gave to me as a child. Fiach, Oona, and Rónán, thank you for suffering endlessly and allowing me to hone my argumentative streak.

Finally, I want to thank my amazing partner. Tanya, I am extremely pleased with the ideas presented in this thesis. Nevertheless, meeting you was undoubtedly the most delightful result. Thank you for all of your help and support over the last three years. Hopefully I can return the favour in the years to come.

## CONTENTS

1 Introduction ..... 1
1.1 Bivariate Information Decomposition ..... 1
1.2 Objectives ..... 3
1.3 Contributions of this Thesis ..... 3
1.4 References ..... 5
2 Background ..... 6
2.1 Information Theory ..... 6
2.2 Analogies Between Entropies and Sets ..... 9
2.3 Lattice Theory ..... 11
2.4 Partial Information Decomposition ..... 14
2.5 Alternative Approaches to Information Decomposition ..... 17
2.6 Summary of Proposed Axioms or Suggested Properties ..... 18
2.7 The Current State of Information Decomposition ..... 20
2.8 References ..... 20
3 Probability Mass Exclusions ..... 24
3.1 Introduction ..... 25
3.2 Information and Eliminations ..... 26
3.3 Information Decomposition and Probability Mass Exclusions ..... 29
3.4 The Directed Components of Mutual Information ..... 29
3.5 Discussion ..... 32
3.A Appendix ..... 33
3.R References ..... 37
4 Pointwise Partial Information Decomposition ..... 39
4.1 Introduction ..... 40
4.1.1 Notation ..... 41
4.1.2 Partial Information Decomposition ..... 42
4.2 Pointwise Information Theory ..... 43
4.2.1 Pointwise Information Decomposition ..... 44
4.2.2 Pointwise Unique ..... 45
4.2.3 Pointwise Partial Information Decomposition ..... 45
4.3 Probability Mass Exclusions and the Directed Components of Pointwise Information ..... 46
4.3.1 Two Distinct Types of Probability Mass Exclusions ..... 47
4.3.2 The Directed Components of Pointwise Information: Specificity and Ambiguity ..... 48
4.3.3 Operational Interpretation of Redundant Information ..... 49
4.3.4 Motivational Example ..... 50
4.4 Pointwise Partial Information Decomposition Using Specificity and Ambiguity ..... 51
4.4.1 Bivariate PPID Using the Specificity and Ambiguity ..... 52
4.4.2 Redundancy Measures on the Specificity and Ambiguity Lattices ..... 53
4.5 Discussion ..... 55
4.5.1 Comparison to Existing Measures ..... 55
4.5.2 Probability Distribution XOR ..... 56
4.5.3 Probability Distribution PWUNQ ..... 57
4.5.4 Probability Distribution RDNERR ..... 58
4.5.5 Probability Distribution TBC ..... 59
4.5.6 Summary of Key Properties ..... 61
4.6 Conclusions ..... 61
4.A Appendix A. Kelly Gambling, Axiom 4, and TBC ..... 62
4.A. 1 Pointwise Side Information and the Kelly Criterion ..... 62
4.A. 2 Justification of Axiom 4 and Redundant Information in TBC ..... 64
4.A. 3 Accumulator Betting and the Target Chain Rule ..... 65
4.B Appendix B. Supporting Proofs and Further Details ..... 66
4.B. 1 Deriving the Specificity and Ambiguity Lattices from Axioms 1-4 ..... 66
4.B. 2 Redundancy Measures on the Lattices ..... 68
4.B. 3 Target Chain Rule ..... 69
4.C Appendix C. Additional Example Probability Distributions ..... 71
4.C. 1 Probability Distribution TBEP ..... 71
4.C. 2 Probability Distribution UNQ ..... 72
4.C. 3 Probability Distribution AND ..... 73
4.R References ..... 73
4.E Erratum ..... 76
5 A New Framework for Information Decomposition ..... 77
5.1 Introduction ..... 78
5.2 Mutual Information Content ..... 80
5.3 Marginal Information Sharing ..... 82
5.4 Synergistic Information Content ..... 85
5.5 Properties of the Union and Intersection Information Content ..... 87
5.6 Generalised Marginal Information Sharing ..... 91
5.7 Multivariate Information Decomposition ..... 92
5.8 Union and Intersection Mutual Information ..... 99
5.9 Conclusion ..... 101
5.A Appendix ..... 104
5.R References ..... 107
5.E Erratum ..... 112
6 Conclusion ..... 113
6.1 Summary of the Main Results ..... 113
6.2 Further Theoretical Research ..... 114
6.3 Potential Applications ..... 115
6.4 References ..... 116
Appendix A Quantifying Information Modification ..... 118
A. 1 Background ..... 119
A. 2 Overview ..... 119
A. 3 Results ..... 120
A.R References ..... 120
Appendix B Implementing the BROJA-Measure in IDTxl ..... 121
B. 1 Summary ..... 122
B.R References ..... 124
Appendix C An Application in Neuroscience ..... 126
C. 1 Introductions ..... 127
C. 2 Methods ..... 129
C.2.1 Definition and Estimation of Unique, Shared and Synergistic Information ..... 129
C.2.2 Mapping of Neural Recordings to Input and Output Variables for PID, and Definition of Information Modification ..... 130
C.2.3 PID Estimation ..... 131
C.2.4 Statistical Testing ..... 131
C.2.5 Electrophysiological Data-Acquisition and Preprocessing ..... 131
C. 3 Results ..... 132
C.3.1 PID of Information Processing in Neural Cultures ..... 132
C. 4 Discussion ..... 134
C.4.1 Which Definition of Synergistic Mutual Information to Use? ..... 135
C.4.2 Previous Studies of Information Modification in Neural Data ..... 136
C.4.3 Information Represented by Multi and Single Unit Data ..... 136
C.4.4 Measured Information Modification versus the Capacity of a Mechanism for Modifying Information ..... 137
C.4.5 On the Distinction between Information Modification and Noise ..... 138
C.4.6 On the Relation between Transfer Entropy and Information Modification ..... 138
C.4.7 New Research Perspectives in Neuroscience Based on PID and Information Modification ..... 139
C. 5 Conclusions ..... 139
C.R References ..... 140

## Chapter 1

## INTRODUCTION

When studying real-world complex systems such as the financial markets or genotype-phenotype mappings, applied researchers frequently seek to determine the strength of the interactions between the various components of the system. Since information-theoretic measures can quantify both linear and non-linear interactions without assuming an underlying model, information theory is a natural choice for this task [1]. Nevertheless, it has recently become clear that the established multivariate information measures can conflate different types of multi-component interdependence [2]. The aim of information decomposition is to develop information-theoretic measures that can separately quantify these distinct modes of dependency.

### 1.1 Bivariate Information Decomposition

Consider a system of three random variables and suppose that we wish to determine how the target variable $T$ depends on the two source variables $S_{1}$ and $S_{2}$. As first discussed by Williams \& Beer [2], there are four distinct modes of dependence that we can consider: the target $T$ could uniquely depend on $S_{1}$, but be independent of $S_{2}$; or vice versa' $T$ could redundantly depend on $S_{1}$ and $S_{2}$ in the same way; or $T$ could be independent of $S_{1}$ and $S_{2}$ individually, but synergistically depend on $S_{1}$ and $S_{2}$ such that together. Figure 1.1 provides three joint probability distributions that each exemplify a single type of dependency.

In general, all four modes of dependency may be present simultaneously, and so the aim is to separately quantify each type of dependence. As such, we require a measure of the unique information $U\left(S_{1} \backslash S_{2} ; T\right)$ from one variable, the unique information $U\left(S_{2} \backslash S_{1} ; T\right)$ from the other variable, the redundant information $R\left(S_{1}, S_{2} ; T\right)$ from either variable, and the synergistic information $C\left(S_{1}, S_{2} ; T\right)$ from both variables. However, Williams \& Beer [2] pointed out that classical information theory provides the following three measures: the mutual information $I\left(S_{1} ; T\right)$, which captures both the unique information from $S_{1}$ and the redundant information between $S_{1}$ and $S_{2}$; the mutual information $I\left(S_{2} ; T\right)$, which captures both the unique information from $S_{2}$ and the same redundant information between $S_{1}$ and $S_{2}$; and the joint mutual information $I\left(S_{1}, S_{2} ; T\right)$ which captures all four types of dependency. Hence, evaluating the decomposition for three variables is fundamentally an algebraic problem whereby we need to solve the following set of equations,

$$
\begin{align*}
I\left(S_{1} ; T\right) & =U\left(S_{1} \backslash S_{2} ; T\right)+R\left(S_{1}, S_{2} ; T\right),  \tag{1.1a}\\
I\left(S_{2} ; T\right) & =U\left(S_{2} \backslash S_{1} ; T\right)+R\left(S_{1}, S_{2} ; T\right),  \tag{1.1b}\\
I\left(S_{1}, S_{2} ; T\right) & =U\left(S_{1} \backslash S_{2} ; T\right)+U\left(S_{2} \backslash S_{1} ; T\right)+R\left(S_{1}, S_{2} ; T\right)+C\left(S_{1}, S_{2} ; T\right) . \tag{1.1c}
\end{align*}
$$

| Unique |  |  |  |
| :--- | :---: | :---: | :---: |
| $\boldsymbol{P}$ | $\boldsymbol{S}_{1}$ | $\boldsymbol{S}_{2}$ | $\boldsymbol{T}$ |
| $1 / 4$ | 0 | 0 | 0 |
| $1 / 4$ | 0 | 1 | 0 |
| $1 / 4$ | 1 | 0 | 1 |
| $1 / 4$ | 1 | 1 | 1 |


$S_{2} O-$


| Synergistic |  |  |  |
| :--- | :--- | :--- | :--- |
| $\boldsymbol{P}$ | $\boldsymbol{S}_{1}$ | $\boldsymbol{S}_{2}$ | $\boldsymbol{T}$ |
| $1 / 4$ | 0 | 0 | 0 |
| $1 / 4$ | 0 | 1 | 1 |
| $1 / 4$ | 1 | 0 | 1 |
| $1 / 4$ | 1 | 1 | 0 |




Figure 1.1: The joint probability distributions given in the three tables here provide an example of the unique information from $S_{1}$ relative to $S_{2}$, redundant information from either $S_{1}$ or $S_{2}$, and synergistic information from both $S_{1}$ and $S_{2}$ together. The circuit diagrams show how one might generate these distributions using binary input sources. For the unique example, the target is a copy of one source, but is independent of the other. In the redundant example, the target is a copy of two correlated sources. Finally, for the synergistic example, the target is given by combining two independent sources using an XOR logic gate such that it is pairwise independent of either source, but is dependent on both sources together. The Venn diagram represents the system of equations (1.1).

From this set of equations, Williams \& Beer [2] showed that the multivariate mutual information conflates redundant dependencies with synergistic dependencies,

$$
\begin{align*}
I\left(S_{1} ; S_{2} ; T\right)= & I\left(S_{1} ; T\right)+I\left(S_{2} ; T\right)-I\left(S_{1}, S_{2} ; T\right) \\
= & U\left(S_{1} \backslash S_{2} ; T\right)+R\left(S_{1}, S_{2} ; T\right)+U\left(S_{2} \backslash S_{1} ; T\right)+R\left(S_{1}, S_{2} ; T\right) \\
& -\left(U\left(S_{1} \backslash S_{2} ; T\right)+U\left(S_{2} \backslash S_{1} ; T\right)+R\left(S_{1}, S_{2} ; T\right)+C\left(S_{1}, S_{2} ; T\right)\right) \\
= & R\left(S_{1}, S_{2} ; T\right)-C\left(S_{1}, S_{2} ; T\right) . \tag{1.2}
\end{align*}
$$

The multivariate mutual information was originally introduced to generalise the notion of mutual information to three or more variables-that is, to quantify the mutual dependence between three or more processes. Clearly, determining the strength of multivariate interactions is an important problem in many areas of science, engineering and economics. However, the multivariate mutual information has the "unfortunate" property that it can be negative [3, p.49]. Prior to the above result, it was not clear what it meant for three or more variables to share negative information, and consequently the multivariate mutual information was said to have "no intuitive meaning" [4].

As we now know thanks to Williams and Beer's result (1.2), the multivariate mutual information is negative when the synergistic information is greater than the redundant information. In order to quantify multivariate dependency, we need to separately quantify each of the distinct kind of dependency. Mathematically, this requires us to solve the set of equations (1.1). However, this set of equations is under-determined and hence solving them requires one more linearly independent equation. Thus, the problem of information decomposition for three variables-that is, quantifying the multivariate dependence between three processes-ultimately comes to providing a suitable definition of either the unique, redundant or synergistic information.

### 1.2 Objectives

The aim of information decomposition is to decompose the information provided by a set of source variables $S_{1}, S_{2}, \ldots, S_{n}$ about a target variable $T$. While it is relatively easy to consider the distinct ways that the target variable can depend on a pair of source variables, it is not immediately obvious how many ways there are for the target variables to depend on an arbitrarily large set of source variables. As such, the goal of information decomposition is to provide the following:

- A framework that accounts for all of the distinct ways in which a target variable can depend on a set of source variables. This framework could be a system of equations, or otherwise.
- A way to quantify each of the distinct modes of dependency.

The partial information decomposition from Williams \& Beer [2] provides a promising candidate for the first of these two requirements. Williams \& Beer also proposed a candidate measure for the second. However, as we will discuss in Section 2.4, this candidate measure was less well-received than their framework.

In response to some of the criticism of the measure from Williams \& Beer [2], Bertschinger et al. [5] suggested that any measure of either unique, redundant or synergistic information should also provide an operational interpretation of the measure. For instance, a principled measure of redundant information should provide a definition of what it means for two or more pieces of information to be considered the same information. Thus, the aim of this thesis is to develop an information decomposition that provides meaningful and interpretable measures of information.

There are two further aims that distinguish this thesis from the existing approaches to information decomposition. Firstly, we want to be able to decompose the information associated with individual realisations. Thus far, our discussion has focused on decomposing the information provided by a set of source variables $S_{1}, S_{2}, \ldots, S_{n}$ about a target variable $T$. However, one can also consider decomposing the pointwise information provided by a set of particular source realisations $s_{1}, s_{2}, \ldots, s_{n}$ about a particular target realisation $t$. Such a pointwise information decomposition would be desirable as it would enable us to localise the information measures so that they can be used to analyse the local information dynamics of time series [6], which would provide far greater detail than merely looking at the average information dynamics of a system.

Secondly, in addition to being able to evaluate the decomposition for any number of sources variables $S_{1}, S_{2}, \ldots, S_{n}$, we wanted the decomposition to produce consistent results for an entire set of target variables $T_{1}, T_{2}, \ldots, T_{k}$ too. This requirement was first suggested by Bertschinger et al. [7] who suggested that the decomposition might satisfy a target chain rule (or left chain rule). Our desire for this property was based upon the importance of the chain rule as a defining characteristic of information.

### 1.3 Contributions of this Thesis

During my candidature, I have made fundamental contributions to the development of the area of information decomposition. This thesis comprises of three papers which were completed during my candidature. The first paper is provided in Chapter 3, and is entitled, "Probability Mass Exclusions and the Directed Components of Mutual Information" [8]. This paper is based upon some of the
often overlooked foundational literature on information theory [9-12]. To be specific, this literature motivated the notion of information by discussing the exclusion of target possibilities induced by individual source values. Intuitively, this idea is exemplified by guessing games such as Twenty Questions or Guess Who?-the more possibilities an inquiry excludes, the greater the amount information you received from that query. This characterisation seemed relevant to the problem of information decomposition, as if different sources exclude the same target possibilities, then these sources must provide the same information. Despite appearing in the early literature, however, this characterisation had never been formalised. This paper contributes a rigorous derivation of information in terms of probability mass exclusions.

Chapter 4 presents the second paper in this thesis, namely the "Pointwise Partial Information Decomposition Using the Specificity and Ambiguity Lattices" [13]. It builds upon the first paper by using the probability mass exclusions to provide a principle way to distinguish between realisations that provide the same information and realisations that merely provide the same amount of information. To be specific, we adopt the following operational interpretation of redundant information: since the pointwise information is ultimately derived from the probability mass exclusions, the same information must induce the same exclusions. Crucially, we use this intuition to derive a pointwise partial information decomposition with a unique set of features relative to other approaches:

- the decomposition can be computed for more than two sources;
- the resultant information measures satisfy a chain rule over targets, which is crucial for producing consistent results regardless of how we choose to perform the analysis; and
- just like the foundational literature, the pointwise decomposition focuses on individual realisations, meaning it can quantify multivariate dependencies between individual realisations, in addition to entire variables.

The final paper contributed during my candidature is provided in Chapter 5, and is entitled, "Generalised Measures of Multivariate Information Content" [14]. This paper provides a bottom-up derivation of what turns out to be an equivalent decomposition to that from the second paper. We begin the paper by asking the following question: if two marginal observers, Alice and Bob, share their information with a third non-observing party, Eve, such that she knows which joint realisation has occurred, and she knows the marginal probability distributions, but she does not know the joint distribution, then how much information does Eve have? We then go on to show that the algebraic structure of shared marginal information is that of a distributive lattice-that is, each distinct way in which a set of marginal observers can share their information with Eve corresponds to an element in a free distributive lattice. We then combine this structure together with the semi-lattice of joint information, and show that the redundancy lattice form partial information decomposition is embedded within this larger algebraic structure. However, since we are considering marginal information contents, this structure is actually equivalent to the specificity lattice from the previous chapter. The paper then closes with a discussion about whether or not one should combine the information contents from the specificity and ambiguity lattices.

### 1.4 References

1. Prokopenko, M., Boschetti, F. \& Ryan, A. J. An information-theoretic primer on complexity, self-organization, and emergence. Complexity 15, 11-28 (2009).
2. Williams, P. L. \& Beer, R. D. Nonnegative decomposition of multivariate information. arXiv preprint arXiv:1004.2515 (2010).
3. Cover, T. M. \& Thomas, J. A. Elements of information theory (John Wiley \& Sons, 2012).
4. Yeung, R. W. A new outlook on Shannon's information measures. IEEE Transactions on Information Theory 37, 466-474. ISSN: 0018-9448 (May 1991).
5. Bertschinger, N., Rauh, J., Olbrich, E., Jost, J. \& Ay, N. Quantifying unique information. Entropy 16, 2161-2183 (2014).
6. Lizier, J. T., Flecker, B. \& Williams, P. L. Towards a synergy-based approach to measuring information modification in Artificial Life (ALIFE), 2013 IEEE Symposium on (2013), 43-51.
7. 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 (2013), 251-269.
8. Finn, C. \& Lizier, J. T. Probability mass exclusions and the directed components of mutual information. Entropy 20, 826 (2018).
9. Hartley, R. V. L. Transmission of information. The Bell System Technical Journal 7, 535-563. ISSN: 0005-8580 (July 1928).
10. Woodward, P. M. \& Davies, I. L. Information theory and inverse probability in telecommunication. Proceedings of the IEE-Part III: Radio and Communication Engineering 99, 37-44 (1952).
11. Fano, R. Transmission of Information. A Statistical Theory of Communications (The MIT Press, 1961).
12. Ash, R. Information Theory ISBN: 9780470034453 (Interscience Publishers, 1965).
13. Finn, C. \& Lizier, J. T. Pointwise partial information decomposition using the specificity and ambiguity lattices. Entropy 20, 297 (2018).
14. Finn, C. \& Lizier, J. T. Generalised measures of multivariate information content. Entropy 22, 216 (2020).

## Chapter 2

## BACKGROUND

Information theory is a branch of probability theory that provides methodologies for quantifying concepts such as uncertainty, surprisal, dependency and information. Notwithstanding some early results from Boltzmann and Gibbs, the roots of modern information theory began with Hartley's derivation of his notion of information [1]. Later, in his seminal "A Mathematical Theory of Communication", Shannon generalised Hartley's definition of information and used this notion to analyse reliability and coding in communication systems [2], which established information theory as a field of study in its own right.

Despite being frequently misunderstood and abused [3] in the decades since its introduction, information theory has become a ubiquitous general tool for quantifying uncertainty and dependency in the sciences. In particular, since information-theoretic measures can quantify both linear and nonlinear interactions without assuming an underlying model, it is particularly well-suited for the analysis of complex systems [4-8]. Nevertheless, as discussed briefly in Chapter 1, it has recently become clear that multivariate information measures can conflate distinct types of multivariate dependencies [9]. We will now introduce some of the fundamental information-theoretic measures and basic ideas from lattice theory, and discuss the various multivariate extensions of information theory.

### 2.1 Information Theory

Let $X$ be a random variable consisting of discrete outcomes $\mathcal{X}$ and consider a particular realisation $x$. The most fundamental function in information theory is the information content of a realisation [10],

$$
\begin{equation*}
h(x)=\log \frac{1}{p(x)}=-\log p(x) . \tag{2.1}
\end{equation*}
$$

(This function is referred to as the pointwise entropy in Chapter 4.) By replacing the probability $p(x)$ with the joint probability $p(x, y)$ or conditional probability $p(x \mid y)$, we can define the joint information content $h(x, y)$ and conditional information content $h(x \mid y)$ respectively. Using the chain rule of probability, it is easy to verify that

$$
\begin{equation*}
h(x, y)=h(x)+h(y \mid x) . \tag{2.2}
\end{equation*}
$$

The information content $h(x)$ satisfies several properties that supports its interpretation as being a measure of information. Indeed, Ash [11] showed that these three characteristics uniquely determine the functional form of the information content. The first property is that if we were already fully
certain that $x$ would occur, then discovering that $x$ occurred provides us with no further information. The function $h$ matches this intuition since $h(x)$ is equal to zero if and only if $p(x)$ is equal to one. The second property is that the more surprising the event $x$, the more information it should convey. Again, clearly the function $h$ satisfies this criteria since as $p(x)$ decreases, the value of $h(x)$ increases monotonically. Finally, if the realisations $x$ and $y$ occur independently, then we would expect that the information provided by knowing that the joint realisation $(x, y)$ occurred is equal to the sum of the information provided by each realisation. Indeed, the function $h(x, y)$ is equal to $h(x)+h(y)$ if and only if $p(x, y)$ is equal to $p(x) p(y)$, i.e. the realisations $x$ and $y$ occur independently.

There is a certain duality between the concepts of information and surprise: if we do not yet know which realisation has occurred, then $h(x)$ quantifies how surprised we would be if the realisation $x$ occurs; on the other hand, if we do know which realisation has occurred, then $h(x)$ quantifies the information associated with this knowledge. In information theory, the notion of information is synonymous with a reduction in surprise.

Perhaps the most well-known function in information theory is the entropy of a random variable, and it is given by the expectation value of the information content $h(x)$ taken over all realisations from the random variable $X$,

$$
\begin{equation*}
H(X)=-\sum_{x \in \mathcal{X}} p(x) \log p(x)=\mathrm{E}_{X}[h(x)] \tag{2.3}
\end{equation*}
$$

Similar to the information content, we can define the joint entropy $H(X, Y)$ and conditional entropy $H(X \mid Y)$. Likewise, these functions are also related to each other by the chain rule,

$$
\begin{equation*}
H(X, Y)=H(X)+H(Y \mid X) \tag{2.4}
\end{equation*}
$$

The duality between the concepts of information and surprise is inherited by the entropy. On one hand, the entropy quantifies the expected surprise of realisations from $X$, or equivalently, our uncertainty before we observe a particular outcome. On the other hand, the entropy quantifies the average amount of information that we expect to get from observing an outcome of the random variable $X$. Again, in information theory, the concept of information can be thought of as the resolution of uncertainty.

The mutual information content or pointwise mutual information is given by

$$
\begin{equation*}
i(x ; y)=h(x)+h(y)-h(x, y)=\log \frac{p(x, y)}{p(x) p(y)} . \tag{2.5}
\end{equation*}
$$

Similar to the information content, we can define the joint mutual information content $i(x, y ; z)$ and conditional mutual information content $i(x, y \mid z)$ by replacing the relevant probabilities with joint or conditional probabilities. This function is perhaps best understood as a comparison between information content $h(x, y)$ associated with the joint realisation $(x, y)$ and the information content $h(x)+h(y)$ associated with $x$ and $y$ occurring independently. However, it is important to note that there is nothing to suggest $h(x, y)$ should be less than $h(x)+h(y)$, and hence the mutual information content $i(x ; y)$ is not non-negative. If the joint event $(x, y)$ is more surprising than both of the marginal events $x$ and $y$ occurring independently, then the mutual information content is negative. Conversely, if the joint event $(x, y)$ is less surprising than both of the marginal events $x$ and $y$ occurring independently, then the information content is positive.


Figure 2.1: A Venn diagram can be used to accurately represent the relationship between the entropies $H(X)$ and $H(Y)$, the joint entropy $H(X, Y)$, the conditional entropies $H(X \mid Y)$ and $H(Y \mid X)$, and the mutual information $I(X ; Y)$.

Similar to how the entropy is equal to the expected information content, the mutual information between two variables is given by the expectation value of the mutual information content,

$$
\begin{align*}
I(X ; Y) & =\mathrm{E}_{(X, Y)}[i(x ; y)] \\
& =\mathrm{E}_{(X, Y)}[h(x)+h(y)-h(x, y)] \\
& =H(X)+H(Y)-H(X, Y) \tag{2.6}
\end{align*}
$$

Similar to the other measures, we can also define the joint mutual information $I(X, Y ; Z)$ and conditional mutual information $I(X, Y \mid Z)$. Despite the fact that the mutual information content is not non-negative, it can be shown that the mutual information is non-negative. Moreover, the mutual information is zero if and only if $X$ and $Y$ are independent [10, 12]. Formally, this follows from the fact that the mutual information can be written as a Kullback-Leibler divergence which is non-negative by Jensen's inequality [12]. Perhaps more intuitively, the information we expect to get from knowing both $X$ and $Y$ simultaneously, i.e. the joint entropy $H(X, Y)$ is upper bounded by the sum of the information we expect to get from knowing $X$ and $Y$ independently. In this way, the mutual information can be interpreted as a measure of dependency, or information that is common to both $X$ and $Y$. Dually, it can also be considered to represent the average reduction in the uncertainty of $X$ that one gets from knowing $Y$, or vice versa. Figure 2.1 shows how this relationship can be represented using a Venn diagram.

Perhaps surprisingly, the vast majority of work in information theory only considers the case whereby a single variable provides information about another variable. Historically, communications theory has been the domain that has driven the development of information theory. When communicating, we are inducing a pairwise relationship between sent messages and received signals, where ideally there is a one-to-one correspondence between the two. With this context in mind, it is perhaps not so surprising that the established measures can conflate the distinct types of multivariate dependency.

The multivariate mutual information content was explicitly defined by Fano [13] and for three variables $X, Y$ and $Z$ is given by

$$
\begin{align*}
i(x ; y ; z) & =h(x)+h(y)+h(z)-h(x, y)-h(x, z)-h(y, z)+h(x, y, z) \\
& =\log \frac{p(x, y) p(x, z) p(y, z)}{p(x) p(y) p(z) p(x, y, z)} . \tag{2.7}
\end{align*}
$$

However, Fano's definition of the multivariate mutual information content was pre-dated by McGill's definition of the multivariate mutual information [14] which is equal to the expectation value of the multivariate mutual information,

$$
\begin{align*}
I(X ; Y ; Z) & =\mathrm{E}_{(X, Y, Z)}[i(x ; y ; z)] \\
& =\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} p(x, y, z) \log \frac{p(x, y) p(x, z) p(y, z)}{p(x) p(y) p(z) p(x, y, z)} \\
& =H(X)+H(Y)+H(Z)-H(X, Y)-H(X, Z)-H(Y, Z)+H(X, Y, Z) \tag{2.8}
\end{align*}
$$

The multivariate mutual information is also known as the interaction information (which is McGill's original name for the quantity), co-information [15], and synergy [16]. It is the natural generalisation of the mutual information beyond two variables, and aims to capture the information that is common to a set of variables using a form of the principle of inclusion-exclusion. More specifically, the multivariate mutual information aims to capture the information that is common to a group of variables in the same way that the principle of inclusion-exclusion can be used to determine the number of elements that are commonly held by a group of sets [17]. Indeed, it is straightforward to see how this definition can be extended to define the multivariate mutual information between an arbitrary number of variables.

However, it has long been known that the multivariate mutual information can be negative [13]. This "odd" [15] property led Czisar and Korner to conclude that the multivariate mutual information has "no intuitive meaning" [18]. Cover and Thomas agree with this sentiment stating that "unfortunately [...] there isn't really a notion of mutual information common to three random variables". Consequently, there is no generally accepted method for quantifying multivariate dependencies. It turns out that, as we will see in Section 2.4, the multivariate mutual information can be negative because it conflates two distinct types of multivariate dependencies.

### 2.2 Analogies Between Entropies and Sets

As we saw in Figure 2.1, the relationship between the entropy of a pair of variables can be understood in terms of a Venn diagram. This interpretation is supported by the fact that for any pair of random variables $X$ and $Y$, the entropy $H$ satisfies the following inequality,

$$
\begin{equation*}
H(X)+H(Y) \geq H(X, Y) \geq H(X), H(Y) \geq 0 \tag{2.9}
\end{equation*}
$$

From this inequality, we get that the conditional entropies and mutual information are non-negative,

$$
\begin{align*}
H(X \mid Y) & =H(X, Y)-H(Y) \geq 0  \tag{2.10}\\
H(Y \mid X) & =H(X, Y)-H(X) \geq 0  \tag{2.11}\\
I(X ; Y) & =H(X)+H(Y)-H(X, Y) \geq 0 \tag{2.12}
\end{align*}
$$

This is analogous to the following inequality that is satisfied by a measure $\mu$ on a pair of sets $A$ and $B$,

$$
\begin{equation*}
\mu(A)+\mu(B) \geq \mu(A \cup B) \geq \mu(A), \mu(B) \geq 0, \tag{2.13}
\end{equation*}
$$

Similar to the conditional entropy and mutual information, we have that

$$
\begin{align*}
& \mu(A \backslash B)=\mu(A \cup B)-\mu(B) \geq 0  \tag{2.14}\\
& \mu(B \backslash A)=\mu(A \cup B)-\mu(A) \geq 0,  \tag{2.15}\\
& \mu(A \cap B)=\mu(A)+\mu(B)-\mu(A \cup B) \geq 0 . \tag{2.16}
\end{align*}
$$

This analogy between entropy and measure has been noted by several authors [18-23].
Nevertheless, MacKay [10] notes that this representation is misleading for at least two reasons. Firstly, since the measure on the intersection $\mu(A \cap B)$ is a measure on a set, it gives the false impression that the mutual information $I(X ; Y)$ is the entropy of some intersection between the random variables. Secondly, it might lead one to believe that this analogy can be generalised beyond two variables. However, as we just saw, the multivariate mutual information is not non-negative, and hence is not analogous to measure. Consequently, it is somewhat dubious to use a Venn diagram to depict the relationship between the entropies of three or more variables, as per Figure 2.2.

Despite the fact that the analogy between entropy and measure is not valid for three or more variables, Yeung [23] showed that there is an analogy between entropy and signed measure that is valid for an arbitrary number of random variables. To do this, Yeung defined a signed measure on a suitably constructed sigma-field that is uniquely determined by the joint entropies of the random variables involved. This correspondence enables one to establish information-theoretic identities from measure-theoretic identities. Thus, Venn diagrams can be used to represent the entropy of three or more variables provided one is aware that the certain overlapping areas may correspond to negative quantities. Nevertheless, a significant issue remains-the multivariate mutual information has "no intuitive meaning" [18, 23].


Figure 2.2: Representing the relationship between three variables using a Venn diagram is potentially misleading since it gives the false impression that the multivariate mutual information $I(X ; Y ; Z)$ is not non-negative, i.e. the hatch area represents a potentially negative quantity.

### 2.3 Lattice Theory

Before introducing the partial information decomposition, we must first discuss some of the basic ideas, definitions and theorems from lattice theory [24-26]. Lattices are themselves a concept from the branch of mathematics known as order theory which describes the notion of order using binary relations and provides a formal framework for describing intuitive statements such as "this object is great than that object" or "this item succeeds this other item". Thus, we will begin by first considering some of the basic definitions from order theory.

Definition 1. A pair $\langle X, \leq\rangle$ is a partially ordered set (or poset) if $\leq$ is a binary relation on a set $X$ such that, for all $a, b, c \in X$, we have
(i) $a \leq a$,
(ii) $a \leq b$ and $b \leq a$ implies $a=b$,
(iii) $a \leq b$ and $b \leq c$ implies $a \leq c$.

These conditions are referred to as reflexivity, anti-symmetry and transitivity, respectively.
Definition 2. A partially ordered set $\langle X, \leq\rangle$ that satisfies the connex property, i.e. is comparable for all $a, b \in X$ such that either $a \leq b$ or $b \leq a$, is called a chain (or a totally ordered set, fully ordered set or linearly ordered set). An antichain is a partially ordered set $\langle X, \leq\rangle$ in which there are no comparable elements.

Definition 3. Let $\langle X, \leq\rangle$ be a partially ordered set, and let $a, b \in X$. We say that $a$ is covered by $b$ (or $b$ covers $a$ ) if $a<b$ and $a \leq c<b$ implies $c=a$. That is, there is no element $c$ from $X$ which satisfies $a<c<b$. The set of elements from $X$ that are covered by $b$ is be denoted $b^{-}$.

Given any partially ordered set $\langle X, \leq\rangle$, we can form a dual partially ordered set $\langle X, \geq\rangle$ by defining $a \geq b$ such that, for every $a, b \in X$, it holds in the latter if and only if $b \leq a$ holds in the former. Crucially, for every statement about the partially ordered set $\langle X, \leq\rangle$, there exists a corresponding dual statement about $\langle X, \geq\rangle$. In general, these dual statements can be found be directly substituting $\leq$ with $\geq$ in the original statement.

Definition 4. We say that a partially ordered set $\langle X, \leq\rangle$ has a bottom element if there exists $\perp \in X$ with the property that $\perp \leq a$ for all $a \in X$. Dually, $X$ has a top element if there exists $T \in X$ such that $a \leq \perp$ for all $a \in X$.

Definition 5. Let $\langle X, \leq\rangle$ be a partially ordered set and let $Y \subseteq X$. Then $a \in Y$ is a maximal element if for all $b \in Y$, we have that $a \leq b$ implies that $a=b$. A minimal element is defined dually. We denote the set of maximal and minimal elements of $Y$ respectively by $\bar{Y}$ and $\underline{Y}$.

Definition 6. Let $\langle X, \leq\rangle$ be a partially ordered set and let $x \in X$. The down-set of $x$, denoted $\downarrow x$, consists of all elements $y \in X$ such that $y \leq x$. The strict down-set of $x$, denoted $\downarrow x$, consists of all elements $y \in X$ such that $y<x$. The up-set and strict up-set of $x$ are defined dually, and are denoted $\uparrow x$ and $\uparrow x$, respectively.

Definition 7. Let $\langle X, \leq\rangle$ be a partially ordered set and let $Y \subseteq X$. An element $a \in X$ is an upper bound for $Y$ if for all $b \in Y$, we have that $b \leq a$. A lower bound for $Y$ is defined dually.

Definition 8. Let $\langle X, \leq\rangle$ be a partially ordered set. An element $a \in X$ is the least upper bound or supremum for $Y$, denoted sup $Y$, if $a$ is an upper bound of $Y$ and for all $b \in Y$ and all $c \in X$, we have that $b \leq c$ implies $a \leq c$. The greatest lower bound or infimum for $Y$, denoted $\inf Y$, is defined dually.

Having considered the basic definitions from order theory, we now wish to consider lattice theory. To begin, a lattice is merely a particular kind of partially ordered set in which, for any pair elements, we can find a unique supremum (called the join) and a unique infimum (called the meet).

Definition 9. A partially ordered set $\langle X, \leq\rangle$ is a lattice if and only if both $\sup \{a, b\}$ and $\inf \{a, b\}$ exist for all $a, b \in X$. We then refer to $\sup \{a, b\}$ and $\inf \{a, b\}$ as the join $a \vee b$ and meet $a \wedge b$, respectively. For $Y \subseteq X$, we denote the meet and join of all elements of $Y$ with $\bigvee Y$ and $\bigwedge Y$, respectively.

For example, the real numbers together form a lattice together with the maximum and minimum operators respectively serving as the join and meet operators. Similarly, the family of all subsets of a set $X$ form a lattice together with the union and intersection operators serving as the respective join and meet operators. As we will see later in this thesis, this latter example is of particular importance.

Although we have introduced lattices as a special kind of partially ordered sets, there is an alternative viewpoint that provides much further insight into these mathematical objects. To be specific, we can view a lattice as an algebraic structure $\langle X, \vee, \wedge\rangle$ and then investigate properties of this structure and its operators. In order to show that the order-theoretic definition of a lattice is equivalent to the algebraic deviation, we will consider three theorems. The first is called the Connecting Lemma, which establishes that the $\vee$ and meet $\wedge$ operators are equivalent to the ordering relation $\leq$. The second considers the algebraic properties of these operators. The third theorem proves that these algebraic properties are sufficient for defining the ordering relation $\leq$.

Lemma 1 (The Connecting Lemma). Let $\langle X, \leq\rangle$ be a lattice and let $a, b \in X$. Then the following are statements are equivalent:
(i) $a \leq b$,
(ii) $a \vee b=b$,
(iii) $a \wedge b=a$.

Proof. See Davey \& Priestley [24].
Theorem 1. Let $\langle X, \leq\rangle$ be a lattice and let $a, b, c \in X$. Then $\vee$ and $\wedge$ satisfy the following identities:

$$
\left.\begin{array}{rl}
a & \vee a=a,  \tag{idempotency}\\
a & \wedge a=a, \\
a & \vee b \\
a & =b \vee a, \\
a & \wedge b \\
=b & \wedge a, \\
(a \vee b) & \vee c
\end{array}\right) a \vee(b \vee c), \quad \text { (idempotency) }, \quad \text { (associativity) }
$$

Proof. See Davey \& Priestley [24].

Definition 10. An algebra $\langle X, \vee, \wedge\rangle$ is a called a lattice if, and only if, $X$ is a non-empty set, and $\vee$ and $\wedge$ are binary operations on $X$ that satisfy the idempotent, commutative, associative and absorption identities.

Theorem 2. Let the algebra $\langle X, \vee, \wedge\rangle$ be a lattice. Then, for all $a, b \in X$, we have the following:
(i) We have that $a \vee b=b$ if and only if $a \wedge b=a$.
(ii) We can define an order operator leq on $X$ by $a \leq b$ if $a \vee b=b$.
(iii) Using $\leq$ from (ii), we have that $\langle X, \leq\rangle$ is a lattice in which $a \vee b=\sup \{a, b\}$ and $a \wedge b=\inf \{a, b\}$.

Proof. See Davey \& Priestley [24].
Thus, we can henceforth use either an order-theoretic definition of a lattice $\langle X, \leq\rangle$ or an algebraic based definition $\langle X, \vee, \wedge\rangle$ and even use them interchangeably knowing that the two definitions are equivalent. This is incredibly useful as the algebraic definition allows us to apply all of the concepts and methods of universal algebra to lattices, which provides a more powerful set of techniques than order theory alone. Let us now consider two further types of partially ordered sets that can also be considered as algebras, namely the join- and meet-semilattices.

Definition 11. A join-semilattice is an algebraic structure $\langle X, V\rangle$ consisting of a set $X$ with a binary operation called join $\vee$ that satisfies the idempotent, commutative and associative identities. A meetsemilattice is an algebraic structure $\langle X, \wedge\rangle$ consisting of a set $X$ with a binary operation called meet $\wedge$ that satisfies the idempotent, commutative and associative identities.

Clearly, a lattice is simultaneously both a join- and meet-semilattice. Indeed, a lattice can be considered to be a special case of either, or more specifically, a lattice is a both a a join-semilattice and meet-semilattice that is connected via the absorption identity. It is this connection that distinguishes a lattice from a semilattice. Next we will consider distributive lattices, which are of particular relevance to this thesis.

Definition 12. A lattice $\langle X, \vee, \wedge\rangle$ is called a distributive lattice if it satisfies the following identities:

$$
\begin{aligned}
& a \vee(b \wedge c)=(a \vee b) \wedge(a \vee c) \\
& a \wedge(b \vee c)=(a \wedge b) \vee(a \wedge c)
\end{aligned}
$$

(distributivity)

The two examples considered earlier in this section are both distributive lattices. That is, the real numbers together with the maximum and minimum operators form a distributive lattice since the maximum and minimum operators are idempotent, commutative, associative and distributive, and are connected by absorption. Similarly, the family of all subsets of a set $X$ together with the union and intersection operations form a distributive lattice since the union and intersection operators are also idempotent, commutative, associative and distributive, and are connected by absorption.

The latter of these two examples is actually highly significant. This is based upon the following theorem, known as Birkoff's Representation Theorem [24, 25, 27] or the Fundamental Theorem of Finite Distributive Lattices [26], which shows us that we can represent every distributive lattice using finite sets. This theorem will be crucial to one of the main results from Chapter 5 of this thesis.

Theorem 3 (Fundamental Theorem of Finite Distributive Lattices). A finite distributive lattice $\langle X, \vee, \wedge\rangle$ is isomorphic to a ring of sets, whereby the lattice's meet and join operations correspond to the intersection and union operations.

Proof. See Davey \& Priestley [24].

### 2.4 Partial Information Decomposition

The partial information decomposition of Williams and Beer [9, 28] was introduced to address the problem of multivariate information decomposition. The approach taken is appealing as rather than speculating about the structure of multivariate information, Williams and Beer took a principled axiomatic approach. Their aim is to decompose the information provided about a target variable $T$ by an arbitrarily large set of predictor variables $S$. They begin by considering potentially overlapping subsets of the predictors $S$ called sources. These sources $A_{1}, A_{2}, \ldots, A_{k}$ are elements of the set of all non-empty subsets of $S$, i.e. elements of the set $\mathcal{P}_{1}(S)=\mathcal{P}(S) \backslash \varnothing$ where $\mathcal{P}(S)$ denotes the power set of $S$. Then they examine the various ways these sources might contain the same information. Formally, they introduce three axioms which "any reasonable measure for redundant information [ $I_{\cap}$ ] should fulfil" [29]. These axioms are based upon the intuition that redundancy should be analogous to the set-theoretic notion of intersection (which is commutative, monotonically decreasing and idempotent).

Axiom 1 (Commutativity). Redundant information is invariant under any permutation $\sigma$ of sources,

$$
I_{\cap}\left(\boldsymbol{A}_{1}, \ldots, \boldsymbol{A}_{k} ; T\right)=I_{\cap}\left(\sigma\left(\boldsymbol{A}_{1}\right), \ldots, \sigma\left(\boldsymbol{A}_{k}\right) ; T\right)
$$

Axiom 2 (Monotonicity). Redundant information decreases monotonically as more sources are included,

$$
I_{\cap}\left(A_{1}, \ldots, A_{k-1} ; T\right) \geq I_{\cap}\left(A_{1}, \ldots, A_{k} ; T\right)
$$

with equality if $\boldsymbol{A}_{k} \supseteq \boldsymbol{A}_{i}$ for any $\boldsymbol{A}_{i} \in\left\{\boldsymbol{A}_{1}, \ldots, \boldsymbol{A}_{k-1}\right\}$.
Axiom 3 (Self-redundancy). Redundant information for a single source $A_{i}$ equals the mutual information,

$$
I_{\cap}\left(A_{i} ; T\right)=I\left(A_{i} ; T\right)
$$

Williams and Beer then consider all of the distinct ways in which the sources $A=\left\{A_{1}, \ldots, A_{k}\right\}$ could contribute redundant information. Before now, we have assumed that the redundancy measure can be applied to any collection of sources, i.e. to elements from the set $\mathcal{P}_{1}(A)$ where $\mathcal{P}_{1}$ denotes the power set with the empty set removed. Since the sources are themselves given by the set $\mathcal{P}_{1}(S)$, the redundancy measure $I_{\cap}$ can be applied to any element from the set $\mathcal{P}_{1}\left(\mathcal{P}_{1}(S)\right)$. This is an enormous set. Nevertheless, we can greatly reduce the number of elements by using Axiom 2. In particular, Axiom 2 states that if $\boldsymbol{A}_{i} \subseteq \boldsymbol{A}_{j}$, then

$$
I_{\cap}\left(\boldsymbol{A}_{j}, \boldsymbol{A}_{i}, \ldots ; t\right)=I_{\cap}\left(\boldsymbol{A}_{i}, \ldots ; t\right)
$$

As such, one only needs to consider the collection of sources such that no source is a superset of any other in order,

$$
\begin{equation*}
\mathcal{A}(S)=\left\{\alpha \in \mathcal{P}_{1}\left(\mathcal{P}_{1}(S)\right) \mid \forall A_{i}, A_{j} \in \alpha, A_{i} \not \subset A_{j}\right\} \tag{2.17}
\end{equation*}
$$

This collection captures all the distinct ways in which the sources could provide redundant information.
Williams and Beer then showed that this set of sources $\mathcal{A}(S)$ is naturally structured. Consider two sets of sources $\alpha, \beta \in \mathcal{A}(S)$. If for every source $B \in \beta$ there exists a source $A \in \alpha$ such that $A \subseteq B$,


Figure 2.3: The redundancy lattice induced by the partial order $\preceq$ from (2.18) over the set of sources $\mathcal{A}(S)$ from (2.17). Each node corresponds to the self-redundancy (Axiom 3) of a source event, e.g. $\{1\}$ corresponds to the source event $\left\{\left\{S_{1}\right\}\right\}$, while $\{12,13\}$ corresponds to the source event $\left\{\left\{S_{1}, S_{2}\right\},\left\{S_{1}, S_{3}\right\}\right\}$. Left: The redundancy lattice for two sources $S=\left\{S_{1}, S_{2}\right\}$. Right: The redundancy lattice for three sources $S=\left\{S_{1}, S_{2}, S_{3}\right\}$.
then all of the information shared by $\boldsymbol{B} \in \beta$ must include any redundant information shared by $\boldsymbol{A} \in \alpha$. Hence, a partial order $\preceq$ can be defined over the elements of the domain $\mathcal{A}(S)$ such that any collection of sources precedes another if and only if the latter provides any information that the former provides,

$$
\begin{equation*}
\forall \alpha, \beta \in \mathcal{A}(\boldsymbol{S}),(\alpha \preceq \beta \Longleftrightarrow \forall \boldsymbol{B} \in \beta, \exists \boldsymbol{A} \in \alpha \mid \boldsymbol{A} \subseteq \boldsymbol{B}) . \tag{2.18}
\end{equation*}
$$

Applying this partial ordering $\preceq$ to the elements of the domain $\mathcal{A}(S)$ produces a redundancy lattice. Figure 2.4 depicts this structure for the case of 2 and 3 predictor variables. Each element in this lattice corresponds to a distinct way in which the set of predictors $S$ can contribute information about the target $T$.

The redundancy measure $I_{\cap}$ can be thought of as a cumulative information function which in effect integrates the contribution from each node as one moves up through the nodes of the lattice. To evaluate the unique information contributed by each node in the lattice, we must evaluate the Möbius inverse $[17,30]$ of the function $I_{\cap}$ over the lattice. That is, the partial information contributed by a node $\alpha$ is given by

$$
\begin{equation*}
I_{\cap}(\alpha)=\sum_{\beta \preceq \alpha} I_{\partial}(\beta) \tag{2.19}
\end{equation*}
$$

By rearranging, it is clear that we can calculate the partial information $I_{\partial}(\alpha)$ recursively from the bottom of the lattice,

$$
\begin{equation*}
I_{\partial}(\alpha)=I_{\cap}(\alpha)-\sum_{\beta \prec \alpha} I_{\partial}(\beta) \tag{2.20}
\end{equation*}
$$

This result generated a great deal of excitement. When it is applied to the redundancy lattice generated by two source variables, it yields the intuitive decomposition (1.1) discussed in Chapter 1. However, it can also be applied to the redundancy lattice for 3 or more sources, and so it provides a general framework for multivariate information decomposition.

Nevertheless, just as we saw with (1.1), evaluating these partial information terms requires an additional definition. But unlike with (1.1), this definition must be a measure of redundant information. Moreover, this definition must satisfy Axioms 1-3. As such, to complete the framework, Williams and Beer simultaneously introduced a measure of redundant information called $I_{\min }$ which quantifies redundancy as the minimum information that any source provides about a target event $t$, averaged over all possible events from $T$,

$$
\begin{equation*}
I_{\min }\left(\boldsymbol{A}_{1}, \ldots, \boldsymbol{A}_{k} ; T\right)=\sum_{t} p(t) \min _{i} I\left(T=t ; \boldsymbol{A}_{i}\right) \tag{2.21}
\end{equation*}
$$

The information that any one source provides is given by the specific information, which quantifies the information associated with a particular realisation $t$ from $T$,

$$
\begin{equation*}
I(T=t ; \boldsymbol{A})=\sum_{\boldsymbol{a}} p(\boldsymbol{a} \mid t)\left[\log \frac{1}{p(t)}-\log \frac{1}{p(t \mid \boldsymbol{a})}\right] \tag{2.22}
\end{equation*}
$$

One the important features of the specific information is that it is non-negative. Indeed, this property is crucially important for the definition (2.21) since the minimum operator would be less meaningful for a potentially negative quantity. For example, if the mutual information content (2.5) were used in-place of the specific information, we could have the situation whereby the mutual information content is positive for one source, but is negative for another source. There is little justification for using the minimum of these sources as a measure of redundancy.

Not long after its introduction, however, $I_{\min }$ was criticised for failing to distinguish between "whether different random variables carry the same information or just the same amount of information" [32, p. 269] (see also [31, 33]). This criticism primarily related to the so-called two-bit copy problem given in Figure 2.4. Several authors argued that the information decomposition induced by

| Two bit copy |  |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
| $\boldsymbol{P}$ | $\boldsymbol{S}_{1}$ | $\boldsymbol{S}_{2}$ | $\boldsymbol{T}$ | $T_{\text {copy }}$ | $T_{\text {alt }}$ |
| $1 / 4$ | 0 | 0 | 0 | 00 | 00 |
| $1 / 4$ | 0 | 1 | 1 | 01 | 01 |
| $1 / 4$ | 1 | 0 | 2 | 10 | 11 |
| $1 / 4$ | 1 | 1 | 3 | 11 | 10 |




Figure 2.4: The decomposition provided by $I_{\min }$ suggests that the two-bit copy provides 1 bit of redundant information and 1 bit synergistic information. Many authors contend that is unreasonable since the distribution can be generated by concatenating two independent bits, as demonstrated by $T_{\text {copy }}$. They argue that since "the wires don't even touch" [31, p.167], the answer should be 1 bit of unique information from each source [31-33]. Nevertheless, using causal systems to provide intuition and insights into information decomposition is dubious as many different causal systems can produced the same probability distribution, as demonstrated by $T_{\text {alt }}$. Since information theory should be agnostic to the labels attached to any outcomes, these different causal systems must lead to the same information decomposition. This final point will be discussed in detail in Chapter 4.
$I_{\min }$ for two-bit copy, i.e. that there is 1 bit of redundant information and 1 bit of synergistic information, was unreasonable as the probability distribution associated with this example could be generated by concatenating two independent bits [31-33]. Since it could be generated by concatenated bits, they argued that any reasonable information decomposition should yield 1 bit of unique information from each source. Nevertheless, as is briefly discussed in Figure 2.4, this reasoning based upon underlying causal dynamics is unreliable.

Although the two-bit copy problem attracted the most attention, there were also two further criticisms of $I_{\min }$ that are particularly relevant to this thesis. The first of these critiques was from Lizier et al. [34], who showed that $I_{\min }$ cannot be used to decompose the pointwise information that is provided by a joint source realisation $s_{1}, s_{2}, \ldots, s_{n}$ about a target realisation $t$. Although the lack of a pointwise information decomposition is not widely cited as a major issue, it is nevertheless a fundamental problem since one could just as rewrite Axioms 1-3 in terms of the pointwise mutual information.

The second relevant criticism comes from Bertschinger et al. [32], and relates to how the information decomposition should behave when considering multiple or joint target variables $T=\left(T_{1}, T_{2}\right)$. Specifically, they suggested that measure of redundant information $I_{\cap}$ should satisfy a target chain rule, or to use the terminology from Bertschinger et al. [32], a left chain rule,

$$
\begin{align*}
I_{\cap}\left(A_{1}, \ldots, A_{k} ; T\right) & =I_{\cap}\left(A_{1}, \ldots, A_{k} ;\left(T_{1}, T_{2}\right)\right) \\
& =I_{\cap}\left(A_{1}, \ldots, A_{k} ; T_{1}\right)+I_{\cap}\left(A_{1}, \ldots, A_{k} ; T_{2} \mid T_{1}\right) . \tag{2.23}
\end{align*}
$$

This requirement is a natural generalisation of the chain rule of mutual information, which is one of the defining characteristics of the notion of information in information theory [2, 13]. In effect, the target chain rule mandates that the information decomposition should be consistent as one considers more and more target variables.

### 2.5 Alternative Approaches to Information Decomposition

In response to the issues with $I_{\min }$, several alternative measures of redundant information have been proposed. The first of these was introduced by Harder et al. [33], who defined a measure of redundant information $I_{\text {red }}$ based upon the methods of information geometry. In the same paper, Harder et al. introduced the so-called identity axiom. This axiom aims to directly address the two-bit copy problem by directly mandating that each source provides 1 bit of unique information for the two-bit copy example.

Axiom 4 (Identity). The redundant information provided by two sources $A_{1}$ and $A_{2}$ about the joint target variable $T=\left(\boldsymbol{A}_{1}, \boldsymbol{A}_{2}\right)$ is equal to the mutual information between the two sources,

$$
I_{\cap}\left(\boldsymbol{A}_{1}, \boldsymbol{A}_{2} ; T\right)=I_{\cap}\left(\boldsymbol{A}_{1}, \boldsymbol{A}_{2} ;\left(\boldsymbol{A}_{1}, \boldsymbol{A}_{2}\right)\right)=I\left(\boldsymbol{A}_{1} ; \boldsymbol{A}_{2}\right) .
$$

An entirely separate approach—which also satisfies Axiom 4-was proposed by Bertschinger et al. [35] who defined a measure of unique information UI based upon the notion that if one variable contains unique information then there must be some way to exploit that information in a decision problem. One of the key points in this paper is that their measure of unique information $\widetilde{\text { UII }}$ (or equivalently, their measure of redundant information $\widetilde{\mathrm{SI}}$ ), only depends on the marginal distribu-
tions $P\left(A_{1}, T\right)$ and $P\left(A_{2}, T\right)$. They argue that any sensible approach should satisfy this requirement and mention that many other candidate measures such as $I_{\min }$ and $I_{\text {red }}$ also satisfy this property. In contrast to other approaches, however, Bertschinger et al. show that their measure of synergistic information $\widetilde{\mathrm{CI}}$ is not attainable with only the marginal distributions $P\left(A_{1}, T\right)$ and $P\left(A_{2}, T\right)$, but rather requires knowledge of the full joint distribution $P\left(A_{1}, A_{2}, T\right)$. Indeed, they showed that $\widetilde{\text { UI }}, \widetilde{\text { SI }}$ and $\widetilde{\mathrm{CI}}$ are the unique functions that simultaneously satisfy both this property and the prior property. In this sense, these two properties can be considered defining characteristics of these functions.

Griffith \& Koch [36] simultaneously and independently defined a measure of synergistic information $S_{V K}$ based upon the idea that synergy should quantify how much greater the whole is than the sum of the parts. In particular, they argue that one should first consider a measure of union information which quantifies the information that is available from all of the individual sources, and subtract this from the total information provided by these sources, i.e. the joint mutual information $I\left(\left(\boldsymbol{A}_{1}, \boldsymbol{A}_{2}\right) ; T\right)$. Griffith \& Koch argue that this union information can be found by destroying the joint relationship between the sources and target, whilst maintaining the marginal relationships between each individual source and the target. It turns out that the resultant measure of synergistic information $S_{\mathrm{VK}}$ is equivalent to $\widetilde{\mathrm{CI}}$ from Bertschinger et al. [35]. Of course, with the benefit of hindsight, this is not surprising since the process outlined by Griffith \& Koch [36] is equivalent to the defining characteristics of functions proposed by Bertschinger et al. [35]. Indeed, this independent derivation and subsequent convergence contributed to the relative popularity of this approach to information decomposition.

In addition to these proposed replacements for $I_{\min }$, there is also a substantial body of literature discussing either partial information decomposition, similar attempts to decompose multivariate information, or the problem of information decomposition in general [29, 31-52]. For comprehensive overview of these approaches, see the editorial that from the Entropy special issue on "Information Decomposition of Target Effects from Multi-Source Interactions: Perspectives on Previous, Current and Future Work" [53].

### 2.6 Summary of Proposed Axioms or Suggested Properties

Almost all of the proposed approaches to information decomposition utilise Axioms 1-3 as the basis for their framework since, as shown by Williams \& Beer [9], these axioms lead to the redundancy lattice which provides a rich algebraic structure for interpreting multivariate dependence. To specific, the information measure $I_{\cap}$ associated with each term in the lattice can then be partitioned via a Möbius inversion into the partial information $I_{\partial}$ associated with each distinct type of multivariate dependence. Any approach that uses these Axioms 1-3 can be referred to as a partial information decomposition. One important property that Williams and Beer discuss is the non-negativity of their partial information terms, a property that became known as local positivity [32].

Property 1 (Local Positivity). The partial information terms should be non-negative, i.e. $I_{\partial} \geq 0$.
Indeed, this property is typically regarded as an essential, since the excitement around information decomposition stems mostly from its ability to provide an explanation as to why the multivariate mutual information is not non-negative. Thus, almost all of the existing approaches to partial information decom-
position have this property. To our knowledge, the only existing approach that does not satisfy Property 1 is due to Ince [47], who justifies this by observing that the negativity is interpretable within the framework of pointwise information theory. It is also worth noting that some authors consider the following property which, to our knowledge, is satisfied by all existing approaches to information decomposition.

Property 2 (Global Positivity). The redundant information should be non-negative, i.e. $I_{\cap} \geq 0$.
Thus far, we have discussed the most popular axioms or properties for information decomposition. Nevertheless, there are a several more less commonly considered axioms or properties to be discussed. Bertschinger et al. [32] suggested that the redundant information should not only be commutative over the sources, but should be symmetric over all of its arguments.

Property 3 (Strong symmetry). The redundant information should be invariant for any permutation $\sigma$,

$$
I_{\cap}\left(\boldsymbol{A}_{1}, \ldots, \boldsymbol{A}_{k} ; T\right)=I_{\cap}\left(\sigma\left(\boldsymbol{A}_{1}\right), \ldots, \sigma\left(\boldsymbol{A}_{k}\right) ; \sigma(T)\right)=I_{\cap}\left(\sigma\left(\boldsymbol{A}_{1}\right), \ldots, \sigma\left(\boldsymbol{A}_{k}\right), \sigma(T)\right)
$$

For the two variable case where there is one source $S$ and one target $T$, this property follows immediately from Axiom 3 and the fact that the mutual information is symmetric. As such, Bertschinger et al. [32] suggest that it would be natural to extend this property to cover an arbitrary number of variables. It is worth noting that this property is a relatively strong requirement, as it immediately constrains most of the nodes in the redundancy lattice [32]. To our knowledge, it is not satisfied by any of the proposed measures of redundant information. In our opinion, it is likely that this is too constraining for any information decomposition that can separately quantify unique, redundant, and synergistic dependence.

In addition to strong symmetry, Bertschinger et al. [32] introduced two further weaker properties which both relate to joint target variables. The first of these is based upon the idea that if a set of sources share some redundant information about a target variable $T_{1}$, then they must share at least as much redundant information about a joint target $T=\left(T_{1}, T_{2}\right)$.

Property 4 (Target monotonicity). The redundant information provided by a set of targets should increase monotonically as more target variables are jointly considered,

$$
I_{\cap}\left(A_{1}, \ldots, A_{k} ; T_{1}\right) \leq I_{\cap}\left(A_{1}, \ldots, A_{k} ;\left(T_{1}, T_{2}\right)\right)
$$

Of course, this property follows immediately from Property 3 and Axiom 2. At this point, an astute reader may notice that the intuition behind target monotonicity is similar to the target chain rule from Section 2.4. Indeed, Bertschinger et al. [32] made this exact point when they introduced the target chain rule as their second property regarding joint target variables.

Property 5 (Target chain rule). The redundant information should satisfy a chain rule for target variables,

$$
\begin{aligned}
I_{\cap}\left(A_{1}, \ldots, A_{k} ; T\right) & =I_{\cap}\left(A_{1}, \ldots, A_{k} ;\left(T_{1}, T_{2}\right)\right) \\
& =I_{\cap}\left(A_{1}, \ldots, A_{k} ; T_{1}\right)+I_{\cap}\left(A_{1}, \ldots, A_{k} ; T_{2} \mid T_{1}\right) \\
& =I_{\cap}\left(A_{1}, \ldots, A_{k} ; T_{2}\right)+I_{\cap}\left(A_{1}, \ldots, A_{k} ; T_{1} \mid T_{2}\right)
\end{aligned}
$$

As was mentioned in Section 2.4, this property is a natural generalisation of the chain rule of mutual information, which is one of the defining characteristics of the notion of information in information theory [2, 13]. One interesting observation made by Bertschinger et al. [32] is that Axiom 4 follows immediately from Properties 1 and 5.

Before moving on, it is important to note that certain combinations of these axioms or properties are not possible. Bertschinger et al. [32] proved that it is not possible to satisfy Axioms 2 and 3, and Properties 1 and 3 simultaneously. Similarly, Rauh et al. [39] showed that no measure caon simultaneously satisfy Axioms 1-4 and Property 1. This latter result is particularly important, as it proves that $\widetilde{U I}, S_{\mathrm{VK}}$ and $I_{\text {red }}$ cannot be generalised to deliver a non-negative information decomposition for an arbitrary number of source variables.

### 2.7 The Current State of Information Decomposition

Information theory is frequently used to quantify the dependency between components in real-world complex systems such as the financial markets or genotype-phenotype mappings. Nevertheless, there are certain questions about these multivariate dependencies that are easy to phrase, but are not quantifiable using classical information theory. For example, do two genes hold the same information redundantly about eye-colour, is it synergistically determined by the pair, or is it perhaps uniquely determined by just one of the pair? Information decomposition promises to provide a framework for addressing such questions.

Unfortunately-in spite of a concerted effort by the information theory and complex systems community-no one approach to information decomposition has been universally accepted. All of the existing proposals have some kind of significant shortcomings. Either they do not work for more than two input sources, or they fail to produce interpretable measures in some edge case. Some of these approaches clash with each other and produced contradictory results for the same underlying system. In short, information decomposition is still considered to be an open problem.

One of the most promising areas of application for information decomposition is in neuroscience. Much of the interest stems from the potential for information decomposition to quantify synergistic dependencies. A measure of synergistic interactions could provide a means to reveal the dynamics of how multiple inputs of information are fused during cognitive tasks, and to improve our ability to infer brain networks from neural imaging data. Indeed, several of the current proposals have been applied to various problems in neuroscience [54-60].

### 2.8 References

1. Hartley, R. V. L. Transmission of information. The Bell System Technical Journal 7, 535-563. ISSN: 0005-8580 (July 1928).
2. Shannon, C. E. A Mathematical Theory of Communication. Bell System Technical Journal 27, 379423. ISSN: 1538-7305. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x (1948).
3. Shannon, C. E. The bandwagon. IRE Transactions on Information Theory 2, 3 (1956).
4. Gell-Mann, M. The Quark and the Jaguar: Adventures in the Simple and the Complex ISBN: 0-7167-2725-0 (W. H. Freeman \& Co., New York, NY, USA, 1995).
5. Adami, C. What is complexity? BioEssays 24, 1085-1094 (2002).
6. Lindgren, K. Information theory for complex systems. Lecture Notes, Department of Physical Resource Theory, Chalmers and Göteborg University (2003).
7. Prokopenko, M., Boschetti, F. \& Ryan, A. J. An information-theoretic primer on complexity, self-organization, and emergence. Complexity 15, 11-28 (2009).
8. Mitchell, M. Complexity: A guided tour (Oxford University Press, 2009).
9. Williams, P. L. \& Beer, R. D. Nonnegative decomposition of multivariate information. arXiv preprint arXiv:1004.2515 (2010).
10. MacKay, D. J. Information theory, inference and learning algorithms (Cambridge university press, 2003).
11. Ash, R. Information Theory ISBN: 9780470034453 (Interscience Publishers, 1965).
12. Cover, T. M. \& Thomas, J. A. Elements of information theory (John Wiley \& Sons, 2012).
13. Fano, R. Transmission of Information. A Statistical Theory of Communications (The MIT Press, 1961).
14. McGill, W. Multivariate information transmission. Transactions of the IRE Professional Group on Information Theory 4, 93-111 (1954).
15. Bell, A. J. The co-information lattice in Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA 2003 (2003).
16. Gawne, T. J. \& Richmond, B. J. How independent are the messages carried by adjacent inferior temporal cortical neurons? Journal of Neuroscience 13, 2758-2771 (1993).
17. Rota, G.-C. On the foundations of combinatorial theory I. Theory of Möbius functions. Probability theory and related fields 2, 340-368 (1964).
18. Csiszar, I. \& Körner, J. Information theory: coding theorems for discrete memoryless systems (Academic Press, Inc., 1981).
19. Reza, F. An Introduction to Information Theory (McGraw-Hill, 1961).
20. Kuo Ting, H. On the amount of information. Teoriya Veroyatnostei i ee Primeneniya 7. In Russian; English translation available, 447-455 (1962).
21. Abramson, N. Information theory and coding (McGraw-Hill, 1963).
22. Campbell, L. Entropy as a measure. IEEE Transactions on Information Theory 11, 112-114 (1965).
23. Yeung, R. W. A new outlook on Shannon's information measures. IEEE Transactions on Information Theory 37, 466-474. ISSN: 0018-9448 (May 1991).
24. Davey, B. A. \& Priestley, H. A. Introduction to lattices and order (Cambridge University Press, 2002).
25. Grätzer, G. General lattice theory (Springer Science \& Business Media, 2002).
26. Stanley, R. P. Enumerative Combinatorics (Cambridge University Press, 1997).
27. Birkhoff, G. Lattice theory (American Mathematical Soc., 1940).
28. Williams, P. L. \& Beer, R. D. Decomposing Multivariate Information. Privately communicated. (2010).
29. Olbrich, E., Bertschinger, N. \& Rauh, J. Information decomposition and synergy. Entropy 17, 3501-3517 (2015).
30. Stanley, R. P. Enumerative Combinatorics 2nd ed. (Cambridge University Press, 2012).
31. Griffith, V., Chong, E. K., James, R. G., Ellison, C. J. \& Crutchfield, J. P. Intersection information based on common randomness. Entropy 16, 1985-2000 (2014).
32. 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 (2013), 251-269.
33. Harder, M., Salge, C. \& Polani, D. Bivariate measure of redundant information. Physical Review E 87, 012130 (2013).
34. Lizier, J. T., Flecker, B. \& Williams, P. L. Towards a synergy-based approach to measuring information modification in Artificial Life (ALIFE), 2013 IEEE Symposium on (2013), 43-51.
35. Bertschinger, N., Rauh, J., Olbrich, E., Jost, J. \& Ay, N. Quantifying unique information. Entropy 16, 2161-2183 (2014).
36. Griffith, V. \& Koch, C. English. in Guided Self-Organization: Inception (ed Prokopenko, M.) 159-190 (Springer Berlin Heidelberg, 2014). ISBN: 978-3-642-53733-2.
37. Harder, M. Information driven self-organization of agents and agent collectives PhD thesis (University of Hertfordshire, 2013).
38. Griffith, V. \& Ho, T. Quantifying redundant information in predicting a target random variable. Entropy 17, 4644-4653 (2015).
39. Rauh, J., Bertschinger, N., Olbrich, E. \& Jost, J. Reconsidering unique information: Towards a multivariate information decomposition in 2014 IEEE International Symposium on Information Theory (2014), 2232-2236.
40. Rauh, J., Banerjee, P. K., Olbrich, E., Jost, J. \& Bertschinger, N. On Extractable Shared Information. Entropy 19. ISSN: 1099-4300 (2017).
41. Rauh, J. et al. Coarse-Graining and the Blackwell Order. Entropy 19. ISSN: 1099-4300 (2017).
42. Rauh, J. Secret sharing and shared information. Entropy 19, 601 (2017).
43. Perrone, P. \& Ay, N. Hierarchical Quantification of Synergy in Channels. Frontiers in Robotics and AI 2, 35 (2016).
44. Rosas, F., Ntranos, V., Ellison, C. J., Pollin, S. \& Verhelst, M. Understanding interdependency through complex information sharing. Entropy 18, 38 (2016).
45. Barrett, A. B. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. Physical Review E 91, 052802 (2015).
46. James, R. G. \& Crutchfield, J. P. Multivariate Dependence beyond Shannon Information. Entropy 19. ISSN: 1099-4300. http: //www.mdpi. com/1099-4300/19/10/531 (2017).
47. Ince, R. Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. Entropy 19, 318 (2017).
48. Ince, R. A. The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. arXiv preprint arXiv:1702.01591 (2017).
49. Chicharro, D. \& Panzeri, S. Synergy and Redundancy in Dual Decompositions of Mutual Information Gain and Information Loss. Entropy 19, 71 (2017).
50. Faes, L., Marinazzo, D. \& Stramaglia, S. Multiscale information decomposition: exact computation for multivariate Gaussian processes. Entropy 19, 408 (2017).
51. Pica, G., Piasini, E., Chicharro, D. \& Panzeri, S. Invariant components of synergy, redundancy, and unique information among three variables. Entropy 19, 451 (2017).
52. Makkeh, A., Theis, D. O. \& Vicente, R. Bivariate Partial Information Decomposition: The Optimization Perspective. Entropy 19, 530 (2017).
53. Lizier, J. T., Bertschinger, N., Jost, J. \& Wibral, M. Information Decomposition of Target Effects from Multi-Source Interactions: Perspectives on Previous, Current and Future Work. Entropy 20 (2018).
54. Angelini, L. et al. Redundant variables and Granger causality. Phys. Rev. E 81, 037201 (3 Mar. 2010).
55. Stramaglia, S. et al. Synergetic and redundant information flow detected by unnormalized Granger causality: application to resting state fMRI. IEEE Transactions on Biomedical Engineering 63, 2518-2524 (2016).
56. Ghazi-Zahedi, K., Langer, C. \& Ay, N. Morphological computation: Synergy of body and brain. Entropy 19, 456 (2017).
57. Maity, A. K., Chaudhury, P. \& Banik, S. K. Information theoretical study of cross-talk mediated signal transduction in MAPK pathways. Entropy 19, 469 (2017).
58. Tax, T., Mediano, P. A. \& Shanahan, M. The partial information decomposition of generative neural network models. Entropy 19, 474 (2017).
59. Kay, J. W., Ince, R. A., Dering, B. \& Phillips, W. A. Partial and Entropic Information Decompositions of a Neuronal Modulatory Interaction. Entropy 19, 560 (2017).
60. Wibral, M., Finn, C., Wollstadt, P., Lizier, J. \& Priesemann, V. Quantifying information modification in developing neural networks via partial information decomposition. Entropy 19, 494 (2017).

## Chapter 3

## Probability Mass Exclusions

In order to utilise the partial information decomposition framework, one must first provide a measure of redundant information. When Williams and Beer [8] first presented $I_{\text {min }}$ it quickly drew criticism for failing to differentiate between variables that provided the same information or merely the same amount of information. As such, providing a principled method for distinguishing between these two possibilities is one of the key challenges in information decomposition.

With this issue in mind, we began searching some of the earliest literature in information theory looking for inspiration. Our preference for the older literature was based upon the fact that pointwise information measures had somewhat fallen out of use in much of the later literature. This search led us eventually to work by Hartley [2], Fano [3] and Ash [4], who all motivated the notion of information by first discussing the exclusions or restrictions induced by received signals. Intuitively, this idea is exemplified by guessing games such as Twenty Questions or Guess Who?-the more possibilities an inquiry excludes, the greater the amount information you received from that query. This description seemed like it could be used to provide a principled method for determining when variables are providing the same information-if different realisations exclude or restrict the same parts of the event space, then these realisations must have been providing the same information.

Despite appearing in this early literature, however, nobody had provided a rigorous characterisation of information in terms of these restrictions. The paper presented in this chapter provides a rigorous derivation of information in terms of probability mass exclusions. The key result is that this characterisation leads to a natural decomposition of the potentially negative mutual information content into two non-negative components-the specificity and ambiguity. The specificity quantifies the information that one could gain about the target from knowing a particular source realisation, and is equal to the information content of that source realisation. The ambiguity, on the other hand, quantifies how useful this source information was in hindsight with respect to the particular target realisation that occurred, and is equal to the information content of that source realisation given the target realisation. In contrast to the mutual information content, there is a one-to-one correspondence between the size of the probability mass exclusions and the decomposed mutual information content. This chapter closes by discussing how this isomorphism might be used to provide an information decomposition that can distinguish between whether realisations and variables provide the same information, rather than merely the same amount of information.

Article

# Probability Mass Exclusions and the Directed Components of Mutual Information 

Conor Finn ${ }^{1,2, *(\text { (1) }}$ and Joseph T. Lizier ${ }^{1}$ (iD)<br>1 Complex Systems Research Group and Centre for Complex Systems, Faculty of Engineering \& IT, The University of Sydney, NSW 2006, Australia; joseph.lizier@sydney.edu.au<br>2 CSIRO Data61, Marsfield, NSW 2122, Australia<br>* Correspondence: conor.finn@sydney.edu.au

Received: 26 September 2018; Accepted: 23 October 2018; Published: 28 October 2018


#### Abstract

Information is often described as a reduction of uncertainty associated with a restriction of possible choices. Despite appearing in Hartley's foundational work on information theory, there is a surprising lack of a formal treatment of this interpretation in terms of exclusions. This paper addresses the gap by providing an explicit characterisation of information in terms of probability mass exclusions. It then demonstrates that different exclusions can yield the same amount of information and discusses the insight this provides about how information is shared amongst random variables-lack of progress in this area is a key barrier preventing us from understanding how information is distributed in complex systems. The paper closes by deriving a decomposition of the mutual information which can distinguish between differing exclusions; this provides surprising insight into the nature of directed information.


Keywords: entropy; mutual information; pointwise; directed; information decomposition

## 1. Introduction

In information theory, there is a duality between the concepts entropy and information: entropy is a measure of uncertainty or freedom of choice, whereas information is a measure of reduction of uncertainty (increase in certainty) or restriction of choice. Interestingly, this description of information as a restriction of choice predates even Shannon [1], originating with Hartley [2]:
> "By successive selections a sequence of symbols is brought to the listener's attention. At each selection there are eliminated all of the other symbols which might have been chosen. As the selections proceed more and more possible symbol sequences are eliminated, and we say that the information becomes more precise."

Indeed, this interpretation led Hartley to derive the measure of information associated with a set of equally likely choices, which Shannon later generalised to account for unequally likely choices. Nevertheless, despite being used since the foundation of information theory, there is a surprising lack of a formal characterisation of information in terms of the elimination of choice. Both Fano [3] and Ash [4] motivate the notion of information in this way, but go on to derive the measure without explicit reference to the restriction of choice. More specifically, their motivational examples consider a set of possible choices $\mathcal{X}$ modelled by a random variable $X$. Then in alignment with Hartley's description, they consider information to be something which excludes possible choices $x$, with more eliminations corresponding to greater information; however, this approach does not capture the concept of information in its most general sense since it cannot account for information provided by partial eliminations which merely reduces the likelihood of a choice $x$ from occurring. (Of course, despite motivating the notion of information in this way, both Fano and Ash provide Shannon's
generalised measure of information which can account for unequally likely choices.) Nonetheless, Section 2 of this paper generalises Hartley's interpretation of information by providing a formal characterisation of information in terms of probability mass exclusions.

Our interest in providing a formal interpretation of information in terms of exclusions is driven by a desire to understand how information is distributed in complex systems [5,6]. In particular, we are interested in decomposing the total information provided by a set of source variables about one or more target variables into the following atoms of information: the unique information provided by each individual source variable, the shared information that could be provided by two or more source variables, and the synergistic information which is only available through simultaneous knowledge of two or more variables [7]. This idea was originally proposed by Williams and Beer who also introduced an axiomatic framework for such a decomposition [8]. However, flaws have been identified with a specific detail in their approach regarding "whether different random variables carry the same information or just the same amount of information" [9] (see also [10,11]). With this problem in mind, Section 3 discusses how probability mass exclusions may provide a principled method for determining if variables provide the same information. Based upon this, Section 4 derives an information-theoretic expression which can distinguish between different probability mass exclusions. Finally, Section 5 closes by discussing how this expression could be used to identify when distinct events provide the same information.

## 2. Information and Eliminations

Consider two random variables $X$ and $Y$ with discrete sample spaces $\mathcal{X}$ and $\mathcal{Y}$, and say that we are trying to predict or infer the value of an event $x$ from $X$ using an event $y$ from $Y$ which has occurred jointly. Ideally, there is a one-to-one correspondence between the occurrence of events from $X$ and $Y$ such that an event $x$ can be exactly predicted using an event $y$. However, in most complex systems, the presence of noise or some other such ambiguity means that we typically do not have this ideal correspondence. Nevertheless, when a particular event $y$ is observed, knowledge of the distributions $P(Y)$ and $P(X, Y)$ can be utilised to improve the prediction on average by using the posterior $P(X \mid y)$ in place of the prior $P(X)$. Our goal now is to understand how Hartley's description relates to the notion of conditional probability.

When a particular event $y$ is observed, we know that the complementary event $\bar{y}=\{\mathcal{Y} \backslash y\}$ did not occur. Thus we can consider the joint distribution $P(X, Y)$ and eliminate the probability mass which is associated with this complementary event $\bar{y}$. In other words, we exclude the probability mass $P(X, \bar{y})$ which leaves only the probability mass $P(X, y)$ remaining. The surviving probability mass can then be normalised by dividing by $p(y)$, which, by definition, yields the conditional distribution $P(X \mid y)$. Hence, with this elimination process in mind, consider the following definition:

Definition 1 (Probability Mass Exclusion). A probability mass exclusion induced by the event $y$ from the random variable $Y$ is the probability mass associated with the complementary event $\bar{y}$, i.e., $p(\bar{y})$.

Echoing Hartley's description, it is perhaps tempting to think that the greater the probability mass exclusion $p(\bar{y})$, the greater the information that $y$ provides about $x$; however, this is not true in general. To see this, consider the joint event $x$ from the random variable $X$. Knowing the event $x$ occurred enables us to categorise the probability mass exclusions induced by $y$ into two distinct types: the first is the portion of the probability mass exclusion associated with the complementary event $\bar{x}$, i.e., $p(\bar{x}, \bar{y})$; while the second is the portion of the exclusion associated with the event $x$, i.e., $p(x, \bar{y})$. Before discussing these distinct types of exclusion, consider the conditional probability of $x$ given $y$ written in terms of these two categories,

$$
\begin{equation*}
p(x \mid y)=\frac{p(x)-p(x, \bar{y})}{1-p(x, \bar{y})-p(\bar{x}, \bar{y})} \tag{1}
\end{equation*}
$$

To see why these two types of exclusions are distinct, consider two special cases: The first special case is when the event $y$ induces exclusions which are confined to the probability mass associated with the complementary event $\bar{x}$. This means that the portion of exclusion $p(\bar{x}, \bar{y})$ is non-zero while the portion $p(x, \bar{y})=0$. In this case the posterior $p(x \mid y)$ is larger than the prior $p(x)$ and is an increasing function of the exclusion $p(\bar{x}, \bar{y})$ for a fixed $p(x)$. This can be seen visually in the probability mass diagram at the top of Figure 1 or can be formally demonstrated by inserting $p(x, \bar{y})=0$ into (1). In this case, the mutual information

$$
\begin{equation*}
i(x ; y)=\log \frac{p(x \mid y)}{p(x)} \tag{2}
\end{equation*}
$$

is a strictly positive, increasing function of $p(\bar{x}, \bar{y})$ for a fixed $p(x)$. (Note that this is the mutual information between events rather than the average mutual information between variables; depending on the context, it is also referred to as the the information density, the pointwise mutual information, or the local mutual information.) For this special case, it is indeed true that the greater the probability mass exclusion $p(\bar{y})$, the greater the information $y$ provides about $x$. Hence, we define this type of exclusion as follows:

Definition 2 (Informative Probability Mass Exclusion). For the joint event xy from the random variables $X$ and $Y$, an informative probability mass exclusion induced by the event $y$ is the portion of the probability mass exclusion associated with the complementary event $\bar{x}$, i.e., $p(\bar{x}, \bar{y})$.

The second special case is when the event $y$ induces exclusions which are confined to the probability mass associated with the event $x$. This means that the portion of exclusion $p(\bar{x}, \bar{y})=0$ while the potion $p(x, \bar{y})$ is non-zero. In this case, the posterior $p(x \mid y)$ is smaller than the prior $p(x)$ and is a decreasing function of the exclusion $p(x, \bar{y})$ for a fixed $p(x)$. This can be seen visually in the probability mass diagram in the middle row of Figure 1 or can be formally demonstrated by inserting $p(\bar{x}, \bar{y})=0$ into (1). In this case, the mutual information (2) is a strictly negative, decreasing function of $p(x, \bar{y})$ for fixed $p(x)$. (Although the mutual information is non-negative when averaged across events from both variables, it may be negative between pairs of events.) This second special case demonstrates that it is not true that the greater the probability mass exclusion $p(\bar{y})$, the greater the information $y$ provides about $x$. Hence, we define this type of exclusion as follows:

Definition 3 (Misinformative Probability Mass Exclusion). For the joint event xy from the random variables $X$ and $Y$, a misinformative probability mass exclusion induced by the event $y$ is the portion of the probability mass exclusion associated with the event $x$, i.e., $p(x, \bar{y})$.

Now consider the general case where both informative and misinformative probability mass exclusions are present simultaneously. It is not immediately clear whether the posterior $p(x \mid y)$ is larger or smaller than the prior $p(x)$, as this depends on the relative size of the informative and misinformative exclusions. Indeed, for a fixed prior $p(x)$, we can vary the informative exclusion $p(\bar{x}, \bar{y})$ whilst still maintaining a fixed posterior $p(x \mid y)$ by co-varying the misinformative exclusion $p(x, \bar{y})$ appropriately; specifically by choosing

$$
\begin{equation*}
p(x, \bar{y})=\frac{p(x)-p(x \mid y)(1-p(\bar{x}, \bar{y}))}{1-p(x \mid y)} \tag{3}
\end{equation*}
$$

Although it is not immediately clear whether the posterior $p(x \mid y)$ is larger or smaller than the prior $p(x)$, the general case maintains the same monotonic dependence as the two constituent special cases. Specifically, if we fix $p(x)$ and the misinformative exclusion $p(x, \bar{y})$, then the posterior $p(x \mid y)$ is an increasing function of the informative exclusion $p(\bar{x}, \bar{y})$. On the other hand, if we fix $p(x)$ and the informative exclusion $p(\bar{x}, \bar{y})$, then the posterior $p(x \mid y)$ is a decreasing function of the misinformative exclusion $p(x, \bar{y})$. This can been seen visually in the probability mass diagram at the bottom of Figure 1
or can be formally demonstrated by fixing and varying the appropriate values for each case in (1). Finally, the relationship between the mutual information and the exclusions in this general case can be explored by inserting (1) into (2), which yields

$$
\begin{equation*}
i(x ; y)=\log \frac{1-p(x, \bar{y}) / p(x)}{1-p(x, \bar{y})-p(\bar{x}, \bar{y})} \tag{4}
\end{equation*}
$$

If $p(x)$ and the misinformative exclusion $p(x, \bar{y})$ are fixed, then $i(x ; y)$ is an increasing function of the informative exclusion $p(\bar{x}, \bar{y})$. On the other hand, if $p(x)$ and the informative exclusion $p(\bar{x}, \bar{y})$ are fixed, then $i(x ; y)$ is a decreasing function of the misinformative exclusion $p(x, \bar{y})$. Finally, if both the informative exclusion $p(\bar{x}, \bar{y})$ and misinformative exclusion $p(x, \bar{y})$ are fixed, the $i(x ; y)$ is an increasing function of $p(x)$.

Now that a formal relationship between eliminations and information has been established using probability theory, we return to the motivational question-can this understanding of information in terms of exclusions aid in our understanding of how random variables share information?


Figure 1. In probability mass diagrams, height represents the probability mass of each joint event from $\mathcal{X} \times Y$ which must sum to 1 . The leftmost of the diagrams depicts the joint distribution $P(X, Y)$, while the central diagrams depict the joint distribution after the occurence of the event $y \in \mathcal{Y}$ leads to exclusion of the probability mass associated with the complementary event $\bar{y}$. By convention, vertical and diagonal hatching represent informative and misinformative exclusions, respectively. The rightmost diagrams represent the conditional distribution after the remaining probability mass has been normalised. Top row: A purely informative probability mass exclusion, $p(\bar{x}, \bar{y})>0$ and $p(x, \bar{y})=0$, leading to $p(x \mid y)>p(x)$ and hence $i(x ; y)>0$. Middle row: A purely misinformative probability mass exclusion, $p(\bar{x}, \bar{y})=0$ and $p(x, \bar{y})>0$, leading to $p(x \mid y)<p(x)$ and hence $i(x ; y)<0$. Bottom row: The general case $p(\bar{x}, \bar{y}>0)$ and $p(x, \bar{y})>0$. Whether $p(x \mid y)$ turns out to be greater or less than $p(x)$ depends on the size of both the informative and misinformative exclusions.

## 3. Information Decomposition and Probability Mass Exclusions

Consider the example in Figure 2 where the events $y$ and $z$ each induce different exclusions, both in terms of size and type, and yet provide the same amount of information about the event $x$ since

$$
\begin{equation*}
i(x ; y)=i(x ; z)=\log 4 / 3 \approx 0.415 \text { bit. } \tag{5}
\end{equation*}
$$

The events $y$ and $z$ reduce our uncertainty about $x$ in distinct ways and yet, after making the relevant exclusions, we have the same freedom of choice about $x$. It is our belief that the information provided by $y$ and $z$ should only be deemed to be the same information if they both reduce our uncertainty about $x$ in the same way; we contend that for the events $y$ and $z$ to reduce our uncertainty about $x$ in the same way, they would have to identically restrict our choice, or make the same exclusions with respect to $x$.


Figure 2. Top: probability mass diagram for $\mathcal{X} \times Y$. Bottom: probability mass diagram for $\mathcal{X} \times Z$. Note that the events $y_{1}$ and $z_{1}$ can induce different exclusions in $P(X)$ and yet still yield the same conditional distributions $P\left(X \mid y_{1}\right)=P\left(X \mid z_{1}\right)$ and hence provide the same amount of information $i\left(x_{1} ; y_{1}\right)=i\left(x_{1} ; z_{1}\right)$ about the event $x_{1}$.

What this example demonstrates is that the mutual information does not-and indeed cannot-distinguish between how events provide information about other events. By definition, the mutual information only depends on the prior $p(x)$ and posterior $p(x \mid y)$ probabilities. Although the posterior $p(x \mid y)$ depends on both the informative and misinformative exclusions, there is no one-to-one correspondence between these exclusions and the resultant mutual information. Indeed, as we saw in (3), there is a continuous range of informative and misinformative exclusions which could yield any given value for the mutual information. As such, any information decomposition based upon the mutual information alone could never distinguish between how events provide information in terms of exclusions. Thus the question naturally arises-can we express the exclusions in terms of information-theoretic measures such that there is a one-to-one correspondence between exclusions and the measures? Such an expression could be utilised in an information decomposition which can distinguish between whether events provide the same information or merely the same amount of information.

## 4. The Directed Components of Mutual Information

The mutual information cannot distinguish between events which induce different exclusions because any given value could arise from a whole continuum of possible informative and misinformative exclusions. Hence, consider decomposing the mutual information into two separate information-theoretic components. Motivated by the strictly positive mutual information observed in the purely informative case and the strictly negative mutual information observed in the purely informative case, let us demand that one of the components be positive while the other component is negative.

Postulate 1 (Decomposition). The information provided by y about $x$ can be decomposed into two non-negative components, such that $i(x ; y)=i_{+}(y \rightarrow x)-i_{-}(y \rightarrow x)$.

Furthermore, let us demand that the two components preserve the functional dependencies between the mutual information and the informative and misinformative exclusion observed in (4) for the general case.

Postulate 2 (Monotonicity). The functions $i_{+}(y \rightarrow x)$ and $i_{-}(y \rightarrow x)$ should satisfy the following conditions:

1. For all fixed $p(x, y)$ and $p(x, \bar{y})$, the function $i_{+}(y \rightarrow x)$ is a continuous, increasing function of $p(\bar{x}, \bar{y})$.
2. For all fixed $p(\bar{x}, y)$ and $p(\bar{x}, \bar{y})$, the function $i_{-}(y \rightarrow x)$ is a continuous, increasing function of $p(x, \bar{y})$.
3. For all fixed $p(x, y)$ and $p(\bar{x}, y)$, the functions $i_{+}(y \rightarrow x)$ and $i_{-}(y \rightarrow x)$ are increasing and decreasing functions of $p(\bar{x}, \bar{y})$, respectively.

Before considering the functions which might satisfy Postulates 1 and 2, there are two further observations to be made about probability mass exclusions. The first observation is that an event $x$ could never induce a misinformative exclusion about itself, since the misinformative exclusion $p(x, \bar{x})=0$. Indeed, inserting this result into the self-information in terms of (4) yields the Shannon information content of the event $x$,

$$
\begin{equation*}
i(x ; x)=\log \frac{1-p(x, \bar{x}) / p(x)}{1-p(x, \bar{x})-p(\bar{x}, \bar{x})}=-\log (1-p(\bar{x}, \bar{x}))=-\log p(x)=h(x) \tag{6}
\end{equation*}
$$

Postulate 3 (Self-Information). An event cannot misinform about itself, hence $i_{+}(x \rightarrow x)=i(x ; x)=h(x)$.
The second observation is that the informative and misinformative exclusions exclusions must individually satisfy the chain rule of probability. As shown in Figure 3, there are three equivalent ways to consider the exclusions induced in $P(X)$ by the events $y$ and $z$. Firstly, we could consider the information provided by the joint event $y z$ which excludes the probability mass in $P(X)$ associated with the joint events $y \bar{z}, \bar{y} z$ and $\bar{y} \bar{z}$. Secondly, we could first consider the information provided by $y$ which excludes the probability mass in $P(X)$ associated with the joint events $\bar{y} z$ and $\bar{y} \bar{z}$, and then subsequently consider the information provided by $z$ which excludes the probability mass in $P(X \mid y)$ associated with the joint event $y \bar{z}$. Thirdly, we could first consider the information provided by $z$ which excludes the probability mass in $P(X)$ associated with the joint events $y \bar{z}$ and $\bar{y} \bar{z}$, and then subsequently consider the information provided by $y$ which excludes the probability mass in $P(X \mid z)$ associated with the joint event $\bar{y} z$. Regardless of the chaining, we start with the same $p(x)$ and finish with the same $p(x \mid y z)$.

Postulate 4 (Chain Rule). The functions $i_{+}(y \rightarrow x)$ and $i_{-}(y \rightarrow x)$ satisfy a chain rule; i.e.,

$$
\begin{aligned}
i_{+}(y z \rightarrow x) & =i_{+}(y \rightarrow x)+i_{+}(z \rightarrow x \mid y) \\
& =i_{+}(z \rightarrow x)+i_{+}(y \rightarrow x \mid z) \\
i_{-}(y z \rightarrow x) & =i_{-}(y \rightarrow x)+i_{-}(z \rightarrow x \mid y) \\
& =i_{-}(z \rightarrow x)+i_{-}(y \rightarrow x \mid z)
\end{aligned}
$$

where the conditional notation denotes the same function only with conditional probability as an argument.
Theorem 1. The unique functions satisfying Postulates 1-4 are

$$
\begin{align*}
& i_{+}(y \rightarrow x)=h(y)=-\log p(y)  \tag{7}\\
& i_{-}(y \rightarrow x)=h(y \mid x)=-\log p(y \mid x) \tag{8}
\end{align*}
$$

By rewriting (7) and (8) in terms of probability mass exclusions, it is easy to verify that Theorem 1 satisfies Postulates 1-4. Perhaps unsurprisingly, this yields a decomposed version of (4),

$$
\begin{align*}
& i_{+}(y \rightarrow x)=-\log (1-p(x, \bar{y})-p(\bar{x}, \bar{y}))  \tag{9}\\
& i_{-}(y \rightarrow x)=-\log \left(1-\frac{p(x, \bar{y})}{p(x)}\right) \tag{10}
\end{align*}
$$

Hence, in order to prove Theorem 1 we must demonstrate that (7) and (8) are the unique functions which satisfy Postulates 1-4. This proof is provided in full in Appendix A.


Figure 3. Top: $y$ and $z$ both simultaneously induce probability mass exclusions in $P(X)$ leading directly to $P(X \mid y, z)$. Middle: $y$ could induce exclusions in $P(X)$ yielding $P(X \mid y)$, and then $z$ could induce exclusions in $P(X \mid y)$ leading to $P(X \mid y, z)$. Bottom: the same as the middle, only vice versa in $y$ and $z$.

## 5. Discussion

Theorem 1 answers the question posed at the end of Section 3-although there is no one-to-one correspondence between these exclusions and the mutual information, there is a one-to-one correspondence between exclusions and the decomposition

$$
\begin{align*}
i(x ; y) & =i_{+}(y \rightarrow x)-i_{-}(y \rightarrow x)  \tag{11}\\
& =h(y)-h(y \mid x)
\end{align*}
$$

It is important to note the directed nature of this decomposition-this equation considers the exclusions induced by $y$ with respect to $x$. It is novel that this particular decomposition enables us to uniquely determine the size of the exclusions induced by $y$ with respect to $x$, rather than $i(x ; y)=h(x)-h(x \mid y)$, which would not satisfy Postulate 4 . Indeed, this latter decomposition is more typically associated with the information provided by $y$ about $x$ since it reflects the change from the prior $p(x)$ to the posterior $p(x \mid y)$. Of course, by Theorem 1 this latter decomposition would allow us to uniquely determine the size exclusions induced by $x$ with respect to $y$.

There is another important asymmetry which can be seen from (9) and (10). The negative component $i_{-}(y \rightarrow x)$ depends on the size of only the misinformative exclusion while the positive component $i_{+}(y \rightarrow x)$ depends on the size of both the informative and misinformative exclusions. The positive component depends on the total size of the exclusions induced by $y$ and hence has no functional dependence on $x$. It quantifies the specificity of the event $y$ : the less likely the outcome $y$ is to occur, the greater the total amount of probability mass excluded by $y$ and therefore the greater the potential for $y$ to inform about $x$. On the other hand, the negative component quantifies the ambiguity of $y$ given $x$ : the less likely the outcome $y$ is to coincide with the outcome $x$, the greater the misinformative probability mass exclusion and therefore the greater the potential for $y$ to misinform about $x$. This asymmetry between the components is apparent when considering the two special cases. In the purely informative case where $p(x, \bar{y})=0$, only the positive informational component is non-zero. On the other hand, in the purely misinformative case, both the positive and negative informational components are non-zero, although it is clear that $i_{+}(y \rightarrow x)<i_{-}(y \rightarrow x)$ and hence $i(x ; y)<0$.

Let us now consider how this information-theoretic expression (which has a one-to-one correspondence with exclusion) could be utilised to provide an information decomposition that can distinguish between whether events provide the same information or merely the same amount of information. Recall the example from Section 3 where $y$ and $z$ provide the same amount of information about $x$, and consider this example in terms of the decomposition (11),

$$
\begin{align*}
& i_{+}(y \rightarrow x)=\log _{2} 8 / 3 \text { bit, } \quad i_{-}(y \rightarrow x)=1 \mathrm{bit} \\
& i_{+}(z \rightarrow x)=\log _{2} 4 / 3 \text { bit, } \quad i_{-}(z \rightarrow x)=0 \mathrm{bit} . \tag{12}
\end{align*}
$$

In contrast to the mutual information in (5), the decomposition reflects the different ways $y$ and $z$ provide information through differing exclusions even if they provide the same amount of information. As for how to decompose multivariate information using this decomposition? This is not the subject of this paper-those who are interested in seen an operational definition of shared information based on redundant exclusions should see [12].

Author Contributions: C.F. and J.L. conceived the idea and wrote the manuscript.
Funding: J.L. was supported through the Australian Research Council DECRA grant DE160100630.
Acknowledgments: We thank Mikhail Prokopenko, Nathan Harding, Nils Bertschinger, and Nihat Ay for helpful discussions relating to this manuscript. We especially thank Michael Wibral for some of our earlier discussions regarding information and exclusions. Finally, we would like to thank the anonymous "Reviewer 2" of [12] for their helpful feedback regarding this paper.

Conflicts of Interest: The authors declare no conflict of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

## Appendix A

This section contains the proof of Theorem 1 . Since it is trivial to verify that (7) and (8) satisfy Postulates $1-4$, the proof will focuses on establishing uniqueness. The proof is structured as follows: Lemma A1 considers the functional form required when $p(\bar{x})=0$ and is used in the proof of Lemma A3; Lemmas A2 and A3 consider the purely informative and misinformative special cases respectively; finally, the proof of Theorem 1 brings these two special cases together for the general case.

The proof of Theorem 1 may seem convoluted, however there are two points to be made about this. Firstly, the proof of Lemma A1 is well-known in functional equation theory [13] and is only given for the sake of completeness. (Accepting this substantially reduces the length of the proof.) Secondly, when establishing uniqueness of the two components, we cannot assume that the components share a common base for the logarithm. Specifically, when considering the purely informative case, Lemma A2 shows that the positive component $i_{+}(y \rightarrow x)$ is a logarithm with same base as the logarithm from Postulate 3, denoted as $b$ throughout. On the other hand, considering the purely misinformative case in Lemma A3 demonstrates that the negative component $i_{-}(y \rightarrow x)$ is a logarithm with base $k$ which is greater than or equal to $b$. When combining these in the proof of Theorem 1, it is necessary to show that $k=b$ in order to prove that the components have a common base.

Lemma A1. In the special case where $p(\bar{x})=0$, we have that $i_{+}(y \rightarrow x)=i_{-}(y \rightarrow x)=-\log _{k} p(y)$ with $k \geq b$, where $b$ is the base of the logarithm from Postulate 3.

Proof. That the logarithm is the unique function which satisfies Postulates $2-4$ is well-known in functional equation theory [13]; however, for the sake of completeness the proof is given here in full. Since $p(\bar{x})=0$, we have that $i(x ; y)=0$ and hence by Postulate 1 , that $i_{+}(y \rightarrow x)=i_{-}(y \rightarrow x)$. Furthermore, we also have that $p(y)=1-p(x, \bar{y})$; thus, without loss of generality, we will consider $i_{-}(y \rightarrow x)$ to be a function of $p(y)$ rather than $p(x, \bar{y})$. As such, let $f(m)$ be our candidate function for $i_{-}(y \rightarrow x)$ where $m=1 / p(y)$. First consider the case where $p(x, \bar{y})=0$, such that $m=1$. Postulate 4 demands that $f(1)=f(1 \cdot 1)=f(1)+f(1)$ and hence $f(1)=0$, i.e., if there is no misinformative exclusion, then the negative informational component should be zero.

Now consider the case where $p(x, \bar{y})$ so that $m$ is a positive integer greater than 1 . If $r$ is an arbitrary positive integer, then $2^{r}$ lies somewhere between two powers of $m$, i.e., there exists a positive integer $n$ such that

$$
\begin{equation*}
m^{n} \leq 2^{r}<m^{n+1} \tag{A1}
\end{equation*}
$$

So long as the base $k$ is greater than 1 , the logarithm is a monotonically increasing function, thus

$$
\begin{equation*}
\log _{k} m^{n} \leq \log _{k} 2^{r}<\log _{k} m^{n+1} \tag{A2}
\end{equation*}
$$

or equivalently,

$$
\begin{equation*}
\frac{n}{r} \leq \frac{\log _{k} 2}{\log _{k} m}<\frac{n+1}{r} \tag{A3}
\end{equation*}
$$

By Postulate 2, $f(m)$ is a monotonically increasing function of $m$, hence applying it to (A1) yields

$$
\begin{equation*}
f\left(m^{n}\right) \leq f\left(2^{r}\right)<f\left(m^{n+1}\right) \tag{A4}
\end{equation*}
$$

Note that, by Postulate 4 and mathematical induction, it is trivial to verify that

$$
\begin{equation*}
f\left(m^{n}\right)=n \cdot f(m) \tag{A5}
\end{equation*}
$$

Hence, by (A4) and (A5), we have that

$$
\begin{equation*}
\frac{n}{r} \leq \frac{f(2)}{f(m)}<\frac{n+1}{r} \tag{A6}
\end{equation*}
$$

Now, (A3) and (A6) have the same bounds, hence

$$
\begin{equation*}
\left|\frac{\log _{k} 2}{\log _{k} m}-\frac{f(2)}{f(m)}\right| \leq \frac{1}{r} \tag{A7}
\end{equation*}
$$

Since $m$ is fixed and $r$ is arbitrary, let $r \rightarrow \infty$. Then, by the squeeze theorem, we get that

$$
\begin{equation*}
\frac{\log _{k} 2}{\log _{k} m}=\frac{f(2)}{f(m)} \tag{A8}
\end{equation*}
$$

and hence,

$$
\begin{equation*}
f(m)=\log _{k} m . \tag{A9}
\end{equation*}
$$

Now consider the case where $p(x, \bar{y})$ so that $m$ is a rational number; in particular, let $m=s / r$ where $s$ and $r$ are positive integers. By Postulate 4,

$$
\begin{equation*}
f(s)=f(s / r \cdot r)=f(s / r)+f(r)=f(m)+f(r) \tag{A10}
\end{equation*}
$$

Thus, combining (A9) and (A10), we get that

$$
\begin{equation*}
f(m)=f(s)-f(r)=\log _{k}(s)-\log _{k}(r)=\log _{k} m \tag{A11}
\end{equation*}
$$

Now consider the case where $p(x, \bar{y})$ such that $m$ is a real number. By Postulate 2, the function (A11) is the unique solution, and hence, $i_{+}(y \rightarrow x)=i_{-}(y \rightarrow x)=-\log _{k} p(y)$.

Finally, to show that $k \geq b$, consider an event $z=y$. By Postulate $3, i_{+}(y \rightarrow z)=-\log _{b} p(y)$. Furthermore, since $p(\bar{z}, \bar{y}) \geq p(\bar{x}, \bar{y})=0$, by Postulate $2, i_{+}(y \rightarrow z) \geq i_{+}(y \rightarrow x)$. Thus, $-\log _{b} p(y) \geq$ $-\log _{k} p(y)$, and hence $k \geq b$.


Figure A1. The probability mass diagram associated with (A12). Lemma A2 uses Postulates 3 and 4 to provide a solution for the purely informative case.

Lemma A2. In the purely informative case where $p(x, \bar{y})=0$, we have that $i_{+}(y \rightarrow x)=-\log _{b} p(y)$ and $i_{-}(y \rightarrow x)=0$, where $b$ is the base of the logarithm from Postulate 3.

Proof. Consider an event $z$ such that $x=y z$ and $\bar{x}=\{y \bar{z}, \bar{y} z, \bar{y} \bar{z}\}$. By Postulate 4,

$$
\begin{equation*}
i_{+}(y z \rightarrow x)=i_{+}(y \rightarrow x)+i_{+}(z \rightarrow x \mid y) \tag{A12}
\end{equation*}
$$

as depicted in Figure A1. By Postulate 3, $i_{+}(y z \rightarrow x)=h(x)$ and $i_{+}(z \rightarrow x \mid y)=h(x \mid y)$, where the latter equality follows from the equivalence of the events $x$ and $z$ given $y$. Furthermore, since $p(x, \bar{y})=0$, we have that $p(x, y)=p(x)$, and hence that $p(y \mid x)=1$. Thus, from (A12), we have that

$$
\begin{align*}
i_{+}(y \rightarrow x) & =h(x)-h(x \mid y) \\
& =h(y)-h(y \mid x)  \tag{A13}\\
& =h(y) .
\end{align*}
$$

Finally, by Postulate $1, i_{-}(y \rightarrow x)=0$.
Lemma A3. In the purely misinformative case where $p(\bar{x}, \bar{y})=0$, we have that $i_{+}(y \rightarrow x)=h(y)-h(y \mid x)-$ $\log _{k} p(y \mid x)$ and $i_{-}(y \rightarrow x)=-\log _{k} p(y \mid x)$ with $k \geq b$, where $b$ is the base of the logarithm from Postulate 3.

Proof. Consider an event $z=x$. By Postulate 4,

$$
\begin{align*}
i_{+}(y z \rightarrow x) & =i_{+}(y \rightarrow x)+i_{+}(z \rightarrow x \mid y)  \tag{A14}\\
& =i_{+}(z \rightarrow x)+i_{+}(y \rightarrow x \mid z) \\
i_{-}(y z \rightarrow x) & =i_{-}(y \rightarrow x)+i_{-}(z \rightarrow x \mid y)  \tag{A15}\\
& =i_{-}(z \rightarrow x)+i_{-}(y \rightarrow x \mid z)
\end{align*}
$$

as depicted in Figure A2. Since $z=x$, by Postulate $3, i_{+}(z \rightarrow x)=h(x), i_{-}(z \rightarrow x)=0$, $i_{+}(z \rightarrow x \mid y)=h(x \mid y)$ and $i_{-}(z \rightarrow x \mid y)=0$. Furthermore, since $p(\bar{x} \mid z)=0$, by Lemma A1, $i_{+}(y \rightarrow x \mid z)=i_{-}(y \rightarrow x \mid z)=-\log _{k} p(y \mid z)=-\log _{k} p(y \mid x)$, hence, from (A14) and (A15), we get that

$$
\begin{align*}
i_{+}(y \rightarrow x) & =h(x)-h(x \mid y)-\log _{k} p(y \mid x)  \tag{A16}\\
& =h(y)-h(y \mid x)-\log _{k} p(y \mid x) \\
i_{-}(y \rightarrow x) & =-\log _{k} p(y \mid x) \tag{A17}
\end{align*}
$$

as required.


Figure A2. The diagram corresponding to (A14) and (A15). Lemma A3 uses Postulate 4 and Lemma A1 to provide a solution for the purely misinformative case.

Proof of Theorem 1. In the general case, both $p(\bar{x}, \bar{y})$ and $p(x, \bar{y})$ are non-zero. Consider two events, $u$ and $v$, such that $y=u v, p(x, \bar{u})=0$ and $p(\bar{x}, \bar{v})=0$. By Postulate 4,

$$
\begin{align*}
& i_{+}(y \rightarrow x)=i_{+}(u v \rightarrow x)=i_{+}(u \rightarrow x)+i_{+}(v \rightarrow x \mid u),  \tag{A18}\\
& i_{-}(y \rightarrow x)=i_{-}(u v \rightarrow x)=i_{-}(u \rightarrow x)+i_{-}(v \rightarrow x \mid u), \tag{A19}
\end{align*}
$$

as depicted in Figure A3. Since $p(x, \bar{u})=0$, by Lemma A2, $i_{+}(u \rightarrow x)=h(u)$ and $i_{-}(u \rightarrow x)=0$; furthermore, we also have that $p(x)=p(x, u)$, and hence $p(v \mid x u)=p(u v \mid x)$. In addition, since
$p(\bar{x}, \bar{v} \mid u)=0$, by Lemma A3, we have that $i_{+}(v \rightarrow x \mid u)=h(v \mid u)+h(v \mid x u)-\log _{k} p(v \mid x u)$ and $i_{-}(v \rightarrow x \mid u)=-\log _{k} p(v \mid x u)$ where $k \geq b$. Therefore, by (A18) and (A19),

$$
\begin{align*}
& i_{+}(y \rightarrow x)= h(u)+h(v \mid u)-h(v \mid x u)-\log _{k} p(v \mid x u)  \tag{A20}\\
&= h(y)-h(y \mid x)-\log _{k} p(y \mid x) \\
& i_{-}(y \rightarrow x)= \\
&=-\log _{k} p(v \mid x u)  \tag{A21}\\
&= \log _{k} p(y \mid x)
\end{align*}
$$

Finally, since Postulate 1 requires that $i_{+}(y \rightarrow x) \geq 0$, we have that $h(y)-h(y \mid x)-\log _{k} p(y \mid x) \geq 0$, or equivalently,

$$
\begin{equation*}
\log _{b} p(y) \leq\left(1-\frac{1}{\log _{b} k}\right) \log _{b} p(y \mid x) \tag{A22}
\end{equation*}
$$

This must hold for all $p(y)$ and $p(y \mid x)$, which is only true in general for $b \geq k$. Hence, $k=b$ and therefore

$$
\begin{align*}
i_{+}(y \rightarrow x) & =h(y)-h(y \mid x)-\log _{b} p(y \mid x)  \tag{A23}\\
& =h(y)  \tag{A24}\\
i_{-}(y \rightarrow x) & =-\log _{b} p(y \mid x) \\
& =h(y \mid x) .
\end{align*}
$$



Figure A3. The probability mass diagram associated with (A18) and (A19). Theorem 1 uses Lemmas A2 and A3 to provide a solution to the general case.

Corollary A1. The conditional decomposition of the information provided by $y$ about $x$ given $z$ is given by

$$
\begin{align*}
& i_{+}(y \rightarrow x \mid z)=h(y \mid z)=-\log p(y \mid z)  \tag{A25}\\
& i_{-}(y \rightarrow x \mid z)=h(y \mid x z)=-\log p(y \mid x z) \tag{A26}
\end{align*}
$$

Proof. Follows trivially using conditional distributions.
Corollary A2. The joint decomposition of the information provided by $y$ and $z$ about $x$ is given by

$$
\begin{align*}
& i_{+}(y z \rightarrow x)=h(y z)=-\log p(y z)  \tag{A27}\\
& i_{-}(y z \rightarrow x)=h(y z \mid x)=-\log p(y z \mid x) \tag{A28}
\end{align*}
$$

The joint decomposition of the information provided by $y$ about $x$ and $z$ is given by

$$
\begin{align*}
& i_{+}(y \rightarrow x z)=h(y)=-\log p(y)  \tag{A29}\\
& i_{-}(y \rightarrow x z)=h(y \mid x z)=-\log p(y \mid x z) \tag{A30}
\end{align*}
$$

Proof. Follows trivially using joint distributions.

Corollary A3. We have the following three identities,

$$
\begin{align*}
& i_{+}(y \rightarrow x)=i_{+}(y \rightarrow z)  \tag{A31}\\
& i_{+}(y \rightarrow x \mid z)=i_{-}(y \rightarrow z)  \tag{A32}\\
& i_{-}(y \rightarrow x \mid z)=i_{-}(y \rightarrow x z) \tag{A33}
\end{align*}
$$

Proof. The identity (A31) follows from (7), while (A32) follows from (8) and (A25); finally, (A33) follows from (A26) and (A30).

Finally, it is not true that the components satisfy a target chain rule. That is, in general the following relation $i_{+}(y \rightarrow x z)=i_{+}(y \rightarrow x)+i_{+}(y \rightarrow z \mid x)$ does not hold, nor does $i_{-}(y \rightarrow x z)=i_{-}(y \rightarrow x)+$ $i_{-}(y \rightarrow z \mid x)$. However, the mutual information must satisfy a chain rule over target events. Thus, it is interesting to observe how the target chain rule for mutual information arises in terms of exclusions. The key observation is that the positive informational component provided by $y$ about $z$ given $x$ equals the negative informational component provided by $y$ about $z$, as per (A32).

Corollary A4. The information provided by $y$ about $x$ and $z$ satisfies the following chain rule,

$$
\begin{equation*}
i(y \rightarrow x z)=i(y \rightarrow x)+i(y \rightarrow z \mid x) \tag{A34}
\end{equation*}
$$

Proof. Starting from the joint decomposition (A29) and (A30). By the identities (A31) and (A33), we get that

$$
\begin{align*}
i(y ; x z) & =i_{+}(y \rightarrow x z)-i_{-}(y \rightarrow x z)  \tag{A35}\\
& =i_{+}(y \rightarrow x)-i_{-}(y \rightarrow z \mid x)
\end{align*}
$$

Then, by identity (A32), and recomposition, we get that

$$
\begin{align*}
i(y ; x z)= & i_{+}(y \rightarrow x)-i_{-}(y \rightarrow x) \\
& +i_{-}(y \rightarrow x)-i_{-}(y \rightarrow z \mid x), \\
= & i_{+}(y \rightarrow x)-i_{-}(y \rightarrow x) \\
& +i_{+}(y \rightarrow z \mid x)-i_{-}(y \rightarrow z \mid x),  \tag{A36}\\
= & i(y ; x)+i(y ; z \mid x) .
\end{align*}
$$

## References

1. Shannon, C.E. A Mathematical Theory of Communication. Bell Syst. Labs Tech. J. 1948, 27, 379-423. [CrossRef]
2. Hartley, R.V.L. Transmission of information. Bell Syst. Labs Tech. J. 1928, 7, 535-563. [CrossRef]
3. Fano, R. Transmission of Information; The MIT Press: Massachusetts, MA, USA, 1961.
4. Ash, R. Information Theory Interscience Tracts in Pure and Applied Mathematics; Interscience Publishers: Hoboken, NJ, USA, 1965.
5. Lizier, J.T. Computation in Complex Systems. In The Local Information Dynamics of Distributed Computation in Complex Systems; Springer: Berlin/Heidelberg, Germany, 2013; pp. 13-52. [CrossRef]
6. Prokopenko, M.; Boschetti, F.; Ryan, A.J. An information-theoretic primer on complexity, self-organization, and emergence. Complexity 2008, 15, 11-28. [CrossRef]
7. Lizier, J.T.; Bertschinger, N.; Jost, J.; Wibral, M. Information Decomposition of Target Effects from Multi-Source Interactions: Perspectives on Previous, Current and Future Work. Entropy 2018, 20, 307. [CrossRef]
8. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information arXiv 2010, arXiv:1004.2515.
9. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J. Shared information-New insights and problems in decomposing information in complex systems. Mathematics 2012, 251-269.
10. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. Phys. Rev. E 2013, 87, 012130. [CrossRef] [PubMed]
11. Griffith, V.; Koch, C. Quantifying Synergistic Mutual Information. In Guided Self-Organization: Inception; Prokopenko, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 159-190.
12. Finn, C.; Lizier, J.T. Pointwise Partial Information Decomposition Using the Specificity and Ambiguity Lattices. Entropy 2018, 20, 297. [CrossRef]
13. Yuichiro, K. Abstract Methods in Information Theory; World Scientific: Singapore, 2016.
(C) 2018 by the authors. 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/).

## Chapter 4

## Pointwise Partial Information Decomposition

In Chapter 3, we showed how the probability mass exclusion can be described in terms of the information-theoretic measures of specificity and ambiguity, i.e. the two distinct terms in the nonnegative decomposition of the pointwise mutual information. The paper presented in this chapter builds upon this result by using the probability mass exclusions to provide a principled way to distinguish between realisations that provide the same information and realisations that merely provide the same amount of information. To be specific, we adopt the following operational interpretation of redundant information: since the pointwise information is ultimately derived from the probability mass exclusions, the same information must induce the same exclusions.

Crucially, since the specificity and ambiguity are non-negative, we will be able to evaluate the partial information decomposition separately on each component, which yields a separate redundancy lattice for the specificity and the ambiguity. Moreover, since these components are defined for every realisation, we will be able to determine these two decompositions for every joint realisation. Thus, by recombining the specificities and ambiguities, we get a redundancy lattice of pointwise mutual information for each realisation-i.e. a pointwise partial information decomposition. Next, we show how the regular partial information decompostion of the mutual information can be found by taking the expectation value of each partial information term in the pointwise lattice. We then show that this decomposition can be evaluated for an arbitrary number of source variables and satisfies a target chain rule, meaning that it can also be evaluated for an arbitrary number of target variables in a consistent manner. To our knowledge, this is the only existing decomposition that satisfies this property [5].

This ability to evaluate a pointwise information decompostion is almost unique amongst the other proposed information decompositions. The only other approach that provides a pointwise information decomposition is due to Ince [18], who uses a set-theoretic interpretation of pointwise information to define a measure of redundant information $I_{\text {ccs }}$. As was discussed in Chapter 2, this interpretation is problematic since the pointwise mutual information is signed. To circumvent this issue, Ince only uses this interpretation to define a measure of redundant information when the relevant signs match and otherwise defines it to be zero. In contrast, the approach presented in this chapter does not dispose of the set-theoretic intuition in these difficult to interpret situations.

Article

# Pointwise Partial Information Decomposition Using the Specificity and Ambiguity Lattices 

Conor Finn ${ }^{1,2, *}$ (1) and Joseph T. Lizier ${ }^{1}$ (iD)<br>1 Complex Systems Research Group and Centre for Complex Systems, Faculty of Engineering \& IT, The University of Sydney, NSW 2006, Australia; joseph.lizier@sydney.edu.au<br>2 CSIRO Data61, Marsfield NSW 2122, Australia<br>* Correspondence: conor.finn@sydney.edu.au

Received: 10 July 2017; Accepted: 10 April 2018; Published: 18 April 2018


#### Abstract

What are the distinct ways in which a set of predictor variables can provide information about a target variable? When does a variable provide unique information, when do variables share redundant information, and when do variables combine synergistically to provide complementary information? The redundancy lattice from the partial information decomposition of Williams and Beer provided a promising glimpse at the answer to these questions. However, this structure was constructed using a much criticised measure of redundant information, and despite sustained research, no completely satisfactory replacement measure has been proposed. In this paper, we take a different approach, applying the axiomatic derivation of the redundancy lattice to a single realisation from a set of discrete variables. To overcome the difficulty associated with signed pointwise mutual information, we apply this decomposition separately to the unsigned entropic components of pointwise mutual information which we refer to as the specificity and ambiguity. This yields a separate redundancy lattice for each component. Then based upon an operational interpretation of redundancy, we define measures of redundant specificity and ambiguity enabling us to evaluate the partial information atoms in each lattice. These atoms can be recombined to yield the sought-after multivariate information decomposition. We apply this framework to canonical examples from the literature and discuss the results and the various properties of the decomposition. In particular, the pointwise decomposition using specificity and ambiguity satisfies a chain rule over target variables, which provides new insights into the so-called two-bit-copy example.


Keywords: mutual information; pointwise information; information decomposition; unique information; redundant information; complementary information; redundancy; synergy

PACS: 89.70.Cf; 89.75.Fb; 05.65.+b; 87.19.lo

## 1. Introduction

The aim of information decomposition is to divide the total amount of information provided by a set of predictor variables, about a target variable, into atoms of partial information contributed either individually or jointly by the various subsets of the predictors. Suppose that we are trying to predict a target variable $T$, with discrete state space $\mathcal{T}$, from a pair of predictor variables $S_{1}$ and $S_{2}$, with discrete state spaces $\mathcal{S}_{1}$ and $\mathcal{S}_{2}$. The mutual information $I\left(S_{1} ; T\right)$ quantifies the information $S_{1}$ individually provides about $T$. Similarly, the mutual information $I\left(S_{2} ; T\right)$ quantifies the information $S_{2}$ individually provides about $T$. Now consider the joint variable $S_{1,2}$ with the state space $\mathcal{S}_{1} \times \mathcal{S}_{2}$. The (joint) mutual information $I\left(S_{1,2} ; T\right)$ quantifies the total information $S_{1}$ and $S_{2}$ together provide about $T$. Although Shannon's information theory provides the prior three measures of information, there are four possible ways $S_{1}$ and $S_{2}$ could contribute information about $T$ : the predictor $S_{1}$ could uniquely provide information about $T$; or the predictor $S_{2}$ could uniquely provide information about $T$;
both $S_{1}$ and $S_{2}$ could both individually, yet redundantly, provide the same information about $T$; or the predictors $S_{1}$ and $S_{2}$ could synergistically provide information about $T$ which is not available in either predictor individually. Thus we have the following underdetermined set of equations,

$$
\begin{align*}
I\left(S_{1,2} ; T\right) & =R\left(S_{1}, S_{2} \rightarrow T\right)+U\left(S_{1} \backslash S_{2} \rightarrow T\right)+U\left(S_{2} \backslash S_{1} \rightarrow T\right)+C\left(S_{1}, S_{2} \rightarrow T\right) \\
I\left(S_{1} ; T\right) & =R\left(S_{1}, S_{2} \rightarrow T\right)+U\left(S_{1} \backslash S_{2} \rightarrow T\right),  \tag{1}\\
I\left(S_{2} ; T\right) & =R\left(S_{1}, S_{2} \rightarrow T\right)+U\left(S_{2} \backslash S_{1} \rightarrow T\right),
\end{align*}
$$

where $U\left(S_{1} \backslash S_{2} \rightarrow T\right)$ and $U\left(S_{2} \backslash S_{1} \rightarrow T\right)$ are the unique information provided by $S_{1}$ and $S_{2}$ respectively, $R\left(S_{1}, S_{2} \rightarrow T\right)$ is the redundant information, and $C\left(S_{1}, S_{2} \rightarrow T\right)$ is the synergistic or complementary information. (The directed notation is utilise here to emphasis the privileged role of the variable T.) Together, the equations in (1) form the bivariate information decomposition. The problem is to define one of the unique, redundant or complementary information-something not provided by Shannon's information theory-in order to uniquely evaluate the decomposition.

Now suppose that we are trying to predict a target variable $T$ from a set of $n$ finite state predictor variables $S=\left\{S_{1}, \ldots, S_{n}\right\}$. In this general case, the aim of information decomposition is to divide the total amount of information $I\left(S_{1}, \ldots, S_{n} ; T\right)$ into atoms of partial information contributed either individually or jointly by the various subsets of $S$. But what are the distinct ways in which these subsets of predictors might contribute information about the target? Multivariate information decomposition is more involved than the bivariate information decomposition because it is not immediately obvious how many atoms of information one needs to consider, nor is it clear how these atoms should relate to each other. Thus the general problem of information decomposition is to provide both a structure for multivariate information which is consistent with the bivariate decomposition, and a way to uniquely evaluate the atoms in this general structure.

In the remainder of Section 1, we will introduce an intriguing framework called partial information decomposition (PID), which aims to address the general problem of information decomposition, and highlight some of the criticisms and weaknesses of this framework. In Section 2, we will consider the underappreciated pointwise nature of information and discuss the relevance of this to the problem of information decomposition. We will then propose a modified pointwise partial information decomposition (PPID), but then quickly repudiate this approach due to complications associated with decomposing the signed pointwise mutual information. In Section 3, we will discuss circumventing this issue by examining information on a more fundamental level, in terms of the unsigned entropic components of pointwise mutual information which we refer to as the specificity and the ambiguity. Then in Section 4-the main section of this paper-we will introduce the PPID using the specificity and ambiguity lattices and the measures of redundancy in Definitions 1 and 2. In Section 5, we will apply this framework to a number of canonical examples from the PID literature, discuss some of the key properties of the decomposition, and compare these to existing approaches to information decomposition. Section 6 will conclude the main body of the paper. Appendix A contains discussions regarding the so-called two-bit-copy problem in terms of Kelly gambling, Appendix B contains many of the technical details and proofs, while Appendix B contains some more examples.

### 1.1. Notation

The following notational conventions are observed throughout this article:

$$
\begin{aligned}
T, \mathcal{T}, t, t^{c} & \text { denote the target variable, event space, event and complementary event respectively; } \\
S, \mathcal{S}, s, s^{c} & \begin{array}{l}
\text { denote the predictor } \text { variable, event space, event and complementary event respectively; } \\
S, s
\end{array} \\
\mathcal{T}^{t}, \mathcal{S}^{s} & \text { represent the set of } n \text { predictor variables }\left\{S_{1}, \ldots, S_{n}\right\} \text { and events }\left\{s_{1}, \ldots, s_{n}\right\} \text { respectively; } \\
H(T), I(S ; T) & \text { uppercase function names be used for average information-theoretic measures; } \\
h(t), i(s, t) & \text { lowercase function names be used for pointwise information-theoretic measures. }
\end{aligned}
$$

When required, the following index conventions are observed:

| $s^{1}, s^{2}, t^{1}, t^{2}$ | superscripts distinguish between different different events in a variable; |
| ---: | :--- |
| $S_{1}, S_{2}, T_{1}, T_{2}$ | subscripts distinguish between different variables; |
| $S_{1,2}, s_{1,2}$ | multiple superscripts represent joint variables and joint events. |

Finally, to be discussed in more detail when appropriate, consider the following:

$$
\begin{aligned}
A_{1}, \ldots, A_{k} & \text { sources are sets of predictor variables, i.e., } A_{i} \in \mathscr{P}_{1}(S) \text { where } \mathscr{P}_{1} \text { is the power set without } \varnothing ; \\
a_{1}, \ldots, a_{k} & \text { source events are sets of predictor events, i.e., } \boldsymbol{a}_{i} \in \mathscr{P}_{1}(\boldsymbol{s}) .
\end{aligned}
$$

### 1.2. Partial Information Decomposition

The partial information decomposition (PID) of Williams and Beer [1,2] was introduced to address the problem of multivariate information decomposition. The approach taken is appealing as rather than speculating about the structure of multivariate information, Williams and Beer took a more principled, axiomatic approach. They start by considering potentially overlapping subsets of $S$ called sources, denoted $A_{1}, \ldots, A_{k}$. To examine the various ways these sources might contain the same information, they introduce three axioms which "any reasonable measure for redundant information [ $I_{\cap}$ ] should fulfil" ([3], p. 3502). Note that the axioms appear explicitly in [2] but are discussed in [1] as mere properties; a published version of the axioms can be found in [4].

W\&B Axiom 1 (Commutativity). Redundant information is invariant under any permutation $\sigma$ of sources,

$$
I_{\cap}\left(A_{1}, \ldots, A_{k} \rightarrow T\right)=I_{\cap}\left(\sigma\left(A_{1}\right), \ldots, \sigma\left(A_{k}\right) \rightarrow T\right)
$$

W\&B Axiom 2 (Monotonicity). Redundant information decreases monotonically as more sources are included,

$$
I_{\cap}\left(A_{1}, \ldots, A_{k-1} \rightarrow T\right) \leq I_{\cap}\left(A_{1}, \ldots, A_{k} \rightarrow T\right)
$$

with equality if $A_{k} \supseteq A_{i}$ for any $A_{i} \in\left\{A_{1}, \ldots, A_{k-1}\right\}$.
W\&B Axiom 3 (Self-redundancy). Redundant information for a single source $\boldsymbol{A}_{i}$ equals the mutual information,

$$
I_{\cap}\left(A_{i} \rightarrow T\right)=I\left(A_{i} ; T\right)
$$

These axioms are based upon the intuition that redundancy should be analogous to the settheoretic notion of intersection (which is commutative, monotonically decreasing and idempotent). Crucially, Axiom 3 ties this notion of redundancy to Shannon's information theory. In addition to these three axioms, there is an (implicit) axiom assumed here known as local positivity [5], which is the requirement that all atoms be non-negative. Williams and Beer [1,2] then show how these axioms reduce the number of sources to the collection of sources such that no source is a superset of any other. These remaining sources are called partial information atoms (PI atoms). Each PI atom corresponds to a distinct way the set of predictors $S$ can contribute information about the target T. Furthermore, Williams and Beer show that these PI atoms are partially ordered and hence form a lattice which they call the redundancy lattice. For the bivariate case, the redundancy lattice recovers the decomposition (1), while in the multivariate case it provides a meaningful structure for decomposition of the total information provided by an arbitrary number of predictor variables.

While the redundancy lattice of PID provides a structure for multivariate information decomposition, it does not uniquely determine the value of the PI atoms in the lattice. To do so requires a definition of a measure of redundant information which satisfies the above axioms. Hence, in order to complete the PID framework, Williams and Beer simultaneously introduced a measure of redundant information called $I_{\min }$ which quantifies redundancy as the minimum information that any source provides about a target event $t$, averaged over all possible events from $T$. However, not long after its introduction $I_{\text {min }}$ was heavily criticised. Firstly, $I_{\min }$ does not distinguish between "whether different random variables carry the same information or just the same amount of information" ([5], p. 269; see also [6,7]). Secondly,
$I_{\min }$ does not possess the target chain rule introduced by Bertschinger et al. [5] (under the name left chain rule). This latter point is problematic as the target chain rule is a natural generalisation of the chain rule of mutual information-i.e., one of the fundamental, and indeed characterising, properties of information in Shannon's theory $[8,9]$.

These issues with $I_{\min }$ prompted much research attempting to find a suitable replacement measure compatible with the PID framework. Using the methods of information geometry, Harder et al. [6] focused on a definition of redundant information called $I_{\text {red }}$ (see also [10]). Bertschinger et al. [11] defined a measure of unique information $\widetilde{U I}$ based upon the notion that if one variable contains unique information then there must be some way to exploit that information in a decision problem. Griffith and Koch [12] used an entirely different motivation to define a measure of synergistic information $\mathcal{S}_{\mathrm{VK}}$ whose decomposition transpired to be equivalent to that of $\widetilde{U I}$ [11]. Despite this effort, none of these proposed measures are entirely satisfactory. Firstly, just as for $I_{\min }$, none of these proposed measures possess the target chain rule. Secondly, these measures are not compatible with the PID framework in general, but rather are only compatible with PID for the special case of bivariate predictors, i.e., the decomposition (1). This is because they all simultaneously satisfy the Williams and Beers axioms, local positivity, and the identity property introduced by Harder et al. [6]. In particular, Rauh et al. [13] proved that no measure satisfying the identity property and the Williams and Beer Axioms 1-3 can yield a non-negative information decomposition beyond the bivariate case of two predictor variables. In addition to these proposed replacements for $I_{\min }$, there is also a substantial body of literature discussing either PID, similar attempts to decompose multivariate information, or the problem of information decomposition in general [3-5,7,10,13-28]. Furthermore, the current proposals have been applied to various problems in neuroscience [29-34]. Nevertheless (to date), there is no generally accepted measure of redundant information that is entirely compatible with PID framework, nor has any other well-accepted multivariate information decomposition emerged.

To summarise the problem, we are seeking a meaningful decomposition of the information provided an arbitrarily large set of predictor variables about a target variable, into atoms of partial information contributed either individually or jointly by the various subsets of the predictors. Crucially, the redundant information must capture when two predictor variables are carrying the same information about the target, not merely the same amount of information. Finally, any proposed measure of redundant information should satisfy the target chain rule so that net redundant information can be consistently computed for consistently for multiple target events.

## 2. Pointwise Information Theory

Both the entropy and mutual information can be derived from first principles as fundamentally pointwise quantities which measure the information content of individual events rather than entire variables. The pointwise entropy $h(t)=-\log p(t)$ quantifies the information content of a single event $t$, while the pointwise mutual information

$$
\begin{equation*}
i(s ; t)=\log \frac{p(t \mid s)}{p(t)}=\log \frac{p(s, t)}{p(s) p(t)}=\log \frac{p(s \mid t)}{p(s)} \tag{2}
\end{equation*}
$$

quantifies the information provided by $s$ about $t$, or vice versa. To our knowledge, these quantities were first considered by Woodward and Davies [35,36] who noted that the average form of Shannon's entropy "tempts one to enquire into other simpler methods of derivation [of the per state entropy]" ([35], p. 51). Indeed, they went on to show that the pointwise entropy and pointwise mutual information can be derived from two axioms concerning the addition of the information provided by the occurrence of individual events [36]. Fano [9] further formalised this pointwise approach by deriving both quantites from four postulates which "should be satisfied by a useful measure of information" ([9], p. 31). Taking the expectation of these pointwise quantities over all events recovers the average entropy $H(T)=\langle h(t)\rangle$ and average mutual information $I(S ; T)=\langle i(s ; t)\rangle$ first derived by Shannon [8]. Although both approaches arrive at the same average quantities, Shannon's treatment
obfuscates the pointwise nature of the fundamental quantities. In contrast, the approach of Woodward, Davis and Fano makes this pointwise nature manifestly obvious.

It is important to note that, in contrast to the average mutual information, the pointwise mutual information is not non-negative. Positive pointwise information corresponds to the predictor event $s$ raising the probability $p(t \mid s)$ relative to the prior probability $p(t)$. Hence when the event $t$ occurs it can be said that the event $s$ was informative about the event $t$. Conversely, negative pointwise information corresponds to the event $s$ lowering the posterior probability $p(t \mid s)$ relative to the prior probability $p(t)$. Hence when the event $t$ occurs we can say that the event $s$ was misinformative about the event $t$. (Not to be confused with disinformation, i.e., intentionally misleading information.) Although a source event $s$ may be misinformative about a particular target event $t$, a source event $s$ is never misinformative about the target variable $T$ since the pointwise mutual information averaged over all target realisations is non-negative [9]. The information provided by $s$ is helpful for predicting $T$ on average; however, in certain instances this (typically helpful) information is misleading in that it lowers $p(t \mid s)$ relative to $p(t)$-typically helpful information which subsequently turns out to be misleading is misinformation.

Finally, before continuing, there are two points to be made about the terminology used to describe pointwise information. Firstly, in certain literature (typically in the context of time-series analysis), the word local is used instead of pointwise, e.g., [4,18]. Secondly, in contemporary information theory, the word average is generally omitted while the pointwise quantities are explicitly prefixed; however, this was not always the accepted convention. Woodward [35] and Fano [9] both referred to pointwise mutual information as the mutual information and then explicitly prefixed the average mutual information. To avoid confusion, we will always prefix both pointwise and average quantities.

### 2.1. Pointwise Information Decomposition

Now that we are familiar with pointwise nature of information, suppose that we have a discrete realisation from the joint event space $\mathcal{T} \times \mathcal{S}_{1} \times \mathcal{S}_{2}$ consisting of the target event $t$ and predictor events $s_{1}$ and $s_{2}$. The pointwise mutual information $i\left(s_{1} ; t\right)$ quantifies the information provided individually by $s_{1}$ about $t$, while the pointwise mutual information $i\left(s_{2} ; t\right)$ quantifies the information provided individually by $s_{2}$ about $t$. The pointwise joint mutual information $i\left(s_{1,2} ; t\right)$ quantifies the total information provided jointly by $s_{1}$ and $s_{2}$ about $t$. In correspondence with the (average) bivariate decomposition (1), consider the pointwise bivariate decomposition, first suggested by Lizier et al. [4],

$$
\begin{align*}
i\left(s_{1,2} ; t\right) & =r\left(s_{1}, s_{2} \rightarrow t\right)+u\left(s_{1} \backslash s_{2} \rightarrow t\right)+u\left(s_{2} \backslash s_{1} \rightarrow t\right)+c\left(s_{1}, s_{2} \rightarrow t\right) \\
i\left(s_{1} ; t\right) & =r\left(s_{1}, s_{2} \rightarrow t\right)+u\left(s_{1} \backslash s_{2} \rightarrow t\right)  \tag{3}\\
i\left(s_{2} ; t\right) & =r\left(s_{1}, s_{2} \rightarrow t\right)+u\left(s_{2} \backslash s_{1} \rightarrow t\right)
\end{align*}
$$

Note that the lower case quantities denote the pointwise equivalent of the corresponding upper case quantities in (1). This decomposition could be considered for every discrete realisation on the support of the joint distribution $P\left(S_{1}, S_{2}, T\right)$. Hence, consider taking the expectation of these pointwise atoms over all discrete realisations,

$$
\begin{align*}
& U\left(S_{1} \backslash S_{2} \rightarrow T\right)=\left\langle u\left(s_{1} \backslash s_{2} \rightarrow t\right)\right\rangle, \quad R\left(S_{1}, S_{2} \rightarrow T\right)=\left\langle r\left(s_{1}, s_{2} \rightarrow t\right)\right\rangle, \\
& U\left(S_{2} \backslash S_{1} \rightarrow T\right)=\left\langle u\left(s_{2} \backslash s_{1} \rightarrow t\right)\right\rangle, \quad C\left(S_{1}, S_{2} \rightarrow T\right)=\left\langle c\left(s_{1}, s_{2} \rightarrow t\right)\right\rangle . \tag{4}
\end{align*}
$$

Since the expectation is a linear operation, this will recover the (average) bivariate decomposition (1). Equation (3) for every discrete realisation, together with (1) and (4) form the bivariate pointwise information decomposition. Just as in (1), these equations are underdetermined requiring a separate definition of either the pointwise unique, redundant or complementary information for uniqueness. (Defining an average atom is sufficient for a unique bivariate decomposition (1), but still leaves the pointwise decomposition (3) within each realisation underdetermined).

### 2.2. Pointwise Unique

Now consider applying this pointwise information decomposition to the probability distribution Pointwise Unique (PWUNQ) in Table 1. In PWUNQ, observing 0 in either of $S_{1}$ or $S_{2}$ provides zero information about the target $T$, while complete information about the outcome of $T$ is obtained by observing 1 or a 2 in either predictor. The probability distribution is structured such that in each of the four realisations, one predictor provides complete information while the other predictor provides zero information-the two predictors never provide the same information about the target which is justified by noting that one of the two predictors always provides zero pointwise information.

Given that redundancy is supposed to capture the same information, it seems reasonable to assume there must be zero pointwise redundant information for each realisation. This assumption is made without any measure of pointwise redundant information; however, no other possibility seems justifiable. This assertion is used to determine the pointwise redundant information terms in Table 1. Then using the pointwise information decomposition (3), we can then evaluate the other pointwise atoms of information in Table 1. Finally using (4), we get that there is zero (average) redundant information, and $1 / 2$ bit of (average) unique information from each predictor. From the pointwise perspective, the only reasonable conclusion seems to be that the predictors in PWUNQ must contain only unique information about the target.

Table 1. Example PWUNQ. For each realisation, the pointwise mutual information provided by each individual and joint predictor events, about the target event has been evaluated. Note that one predictor event always provides full information about the target while the other provides zero information. Based on the this, it is assumed that there must be zero redundant information. The pointwise partial information (PPI) atoms are then calculated via (3).

| $p$ | $s_{1}$ | $s_{2}$ | $t$ | $i\left(s_{1} ; t\right)$ | $i\left(s_{2} ; t\right)$ | $i\left(s_{1,2} ; t\right)$ | $u\left(s_{1} \backslash s_{2} \rightarrow t\right)$ | $u\left(s_{2} \backslash s_{1} \rightarrow t\right)$ | $r\left(s_{1}, s_{2} \rightarrow t\right)$ | $c\left(s_{1}, s_{2} \rightarrow t\right)$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1/4 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| 1/4 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| 1/4 | 0 | 2 | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| 1/4 | 2 | 0 | 2 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| Expected values |  |  |  | $1 / 2$ | 1/2 | 1 | 1/2 | 1/2 | 0 | 0 |

However, in contrast to the above, $I_{\min }, I_{\text {red }}, \widetilde{U I}$, and $\mathcal{S}_{\mathrm{VK}}$ all say that the predictors in PWUNQ contain no unique information, rather only $1 / 2$ bit of redundant information plus $1 / 2$ bit of complementary information. This problem, which will be referred to as the pointwise unique problem, is a consequence of the fact that these measures all satisfy Assumption (*) of Bertschinger et al. [11], which (in effect) states that the unique and redundant information should only depend on the marginal distributions $P\left(S_{1}, T\right)$ and $P\left(S_{2}, T\right)$. In particular, any measure which satisfies Assumption (*) will yield zero unique information when $P\left(S_{1}, T\right)$ is isomorphic to $P\left(S_{2}, T\right)$, as is the case for PWUNQ. (Here, isomorphic should be taken to mean isomorphic probability spaces, e.g., [37], p. 27 or [38], p. 4.) It arises because Assumption (*) (and indeed the operational interpretation the led to its introduction) does not respect the pointwise nature of information. This operational view does not take into account the fact that individual events $s_{1}$ and $s_{2}$ may provide different information about the event $t$, even if the probability distributions $P\left(S_{1}, T\right)$ and $P\left(S_{2}, T\right)$ are the same. Hence, we contend that for any measure to capture the same information (not merely the same amount), it must respect the pointwise nature of information.

### 2.3. Pointwise Partial Information Decomposition

With the pointwise unique problem in mind, consider constructing an information decomposition with the pointwise nature of information as an inherent property. Let $a_{1}, \ldots, a_{k}$ be potentially intersecting subsets of the predictor events $s=\left\{s_{1}, \ldots, s_{n}\right\}$, called source events. Now consider rewriting the Williams and Beer axioms in terms of a measure of pointwise redundant information $i_{\cap}$ where the aim is to deriving a pointwise partial information decomposition (PPID).

PPID Axiom 1 (Symmetry). Pointwise redundant information is invariant under any permutation $\sigma$ of source events,

$$
i_{\cap}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t\right)=i_{\cap}\left(\sigma\left(\boldsymbol{a}_{1}\right), \ldots, \sigma\left(\boldsymbol{a}_{k}\right) \rightarrow T\right)
$$

PPID Axiom 2 (Monotonicity). Pointwise redundant information decreases monotonically as more source events are included,

$$
i_{\cap}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k-1} \rightarrow t\right) \leq i_{\cap}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t\right)
$$

with equality if $\boldsymbol{a}_{k} \supseteq \boldsymbol{a}_{i}$ for any $\boldsymbol{a}_{i} \in\left\{\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k-1}\right\}$.
PPID Axiom 3 (Self-redundancy). Pointwise redundant information for a single source event $\boldsymbol{a}_{i}$ equals the pointwise mutual information,

$$
i_{\cap}\left(\boldsymbol{a}_{i} \rightarrow t\right)=i\left(\boldsymbol{a}_{i} ; t\right)
$$

It seems that the next step should be to define some measure of pointwise redundant information which is compatible with these PPID axioms; however, there is a problem-the pointwise mutual information is not non-negative. While this would not be an issue for the examples like PWUNQ, where none of the source events provide negative pointwise information, it is an issue in general (e.g., see RDNERR in Section 5.4). The problem is that set-theoretic intuition behind Axiom 2 (monotonicity) makes little sense when considering signed measures like the pointwise mutual information.

Given the desire to address the pointwise unique problem, there is a need to overcome this issue. Ince [18] suggested that the set-theoretic intuition is only valid when all source events provide either positive or negative pointwise information. Ince contends that information and misinformation are "fundamentally different" ([18], p. 11) and that the set-theoretic intuition should be admitted in the difficult to interpret situations where both are present. We however, will take a different approach-one which aims to deal with these difficult to interpret situations whilst preserving the set-theoretic intuition that redundancy corresponds to overlapping information.

By way of a preview, we first consider precisely how an event $s_{1}$ provides information about an event $t$ by the means of two distinct types of probability mass exclusion. We show how considering the process in this way naturally splits the pointwise mutual information into particular entropic components, and how one can consider redundancy on each of these components separately. Splitting the signed pointwise mutual information into these unsigned entropic components circumvents the above issue with Axiom 2 (monotonicity). Crucially, however, by deriving these entropic components from the probability mass exclusions, we retain the set-theoretic intuition of redundancy-redundant information will correspond to overlapping probability mass exclusions in the two-event partition $\mathcal{T}^{t}=\left\{t, t^{c}\right\}$.

## 3. Probability Mass Exclusions and the Directed Components of Pointwise Mutual Information

By definition, the pointwise information provided by $s$ about $t$ is associated with a change from the prior $p(t)$ to the posterior $p(t \mid s)$. As we explored from first principles in Finn and Lizier [39], this change is a consequence of the exclusion of probability mass in the target distribution $P(T)$ induced by the occurrence of the event $s$ and inferred via the joint distribution $P(S, T)$. To be specific, when the event $s$ occurs, one knows that the complementary event $s^{c}=\{\mathcal{S} \backslash s\}$ did not occur. Hence one can exclude the probability mass in the joint distribution $P(S, T)$ associated with the complementary event, i.e., exclude $P\left(s^{\mathrm{c}}, T\right)$, leaving just the probability mass $P(s, T)$ remaining. The new target distribution $P(T \mid s)$ is evaluated by normalising this remaining probability mass. In [39] we introduced probability mass diagrams in order to visually explore the exclusion process. Figure 1 provides an example of such a diagram. Clearly, this process is merely a description of the definition of conditional probability. Nevertheless, we content that by viewing the change from the prior to the posterior in this way-by focusing explicitly on the exclusions rather than the resultant conditional probability-the vague intuition that redundancy corresponds to overlapping information becomes more apparent. This point will elaborated upon in Section 3.3. However, in order
to do so, we need to first discuss the two distinct types of probability mass exclusion (which we do in Section 3.1) and then relate these to information-theoretic quantities (which we do in Section 3.2).


Figure 1. Sample probability mass diagrams, which use length to represent the probability mass of each joint event from $\mathcal{T} \times \mathcal{S}$. (Left) the joint distribution $P(S, T)$; (Middle) The occurrence of the event $s^{1}$ leads to exclusions of the complementary event $s^{1 c}$ which consists of two elementary event, i.e., $s^{1^{c}}=\left\{s^{2}, s^{3}\right\}$. This leaves the probability mass $P\left(s^{1}, T\right)$ remaining. The exclusion of the probability mass $p\left(s^{1^{c}}, t^{1}\right)$ was misinformative since the event $t^{1}$ did occur. By convention, misinformative exclusions will be indicated with diagonal hatching. On the other hand, the exclusion of the probability mass $p\left(t^{1^{c}}, s^{1^{c}}\right)$ was informative since the complementary event $t^{1^{c}}$ did not occur. By convention, informative exclusions will be indicated with horizontal or vertical hatching; (Right) this remaining probability mass can be normalised yielding the conditional distribution $P\left(T \mid s^{1}\right)$.

### 3.1. Two Distinct Types of Probability Mass Exclusions

In [39] we examined the two distinct types of probability mass exclusions. The difference between the two depends on where the exclusion occurs in the target distribution $P(T)$ and the particular target event $t$ which occurred. Informative exclusions are those which are confined to the probability mass associated with the set of elementary events in the target distribution which did not occur, i.e., exclusions confined to the probability mass of the complementary event $p\left(t^{c}\right)$. They are called such because the pointwise mutual information $i(s ; t)$ is a monotonically increasing function of the total size of these exclusions $p\left(t^{c}\right)$. By convention, informative exclusions are represented on the probability mass diagrams by horizontal or vertical lines. On the other hand, the misinformative exclusion is confined to the probability mass associated with the elementary event in the target distribution which did occur, i.e., an exclusion confined to $p(t)$. It is referred to as such because the pointwise mutual information $i(s ; t)$ is a monotonically decreasing function of the size of this type of exclusion $p(t)$. By convention, misinformative exclusions are represented on the probability mass diagrams by diagonal lines.

Although an event $s$ may exclusively induce either type of exclusion, in general both types of exclusion are present simultaneously. The distinction between the two types of exclusions leads naturally to the following question-can one decompose the pointwise mutual information $i(s ; t)$ into a positive informational component associated with the informative exclusions, and a negative informational component associated with the misinformative exclusions? This question is considered in detail in Section 3.2. However, before moving on, there is a crucial observation to be made about the pointwise mutual information which will have important implications for the measure of redundant information to be introduced later.

Remark 1. The pointwise mutual information $i(s ; t)$ depends only on the size of informative and misinformative exclusions. In particular, it does not depend on the apportionment of the informative exclusions across the set of elementary events contained in the complementary event $t^{c}$.

In other words, whether the event $s$ turns out to be net informative or misinformative about the event $t$-whether $i(s ; t)$ is positive or negative-depends on the size of the two types of exclusions; but, to be explicit, does not depend on the distribution of the informative exclusion across the set of
target events which did not occur. This remark will be crucially important when it comes to providing the operational interpretation of redundant information in Section 3.3. (It is also further discussed in terms of Kelly gambling [40] in Appendix A).

### 3.2. The Directed Components of Pointwise Information: Specificity and Ambiguity

We return now to the idea that one might be able to decompose the pointwise mutual information into a positive and negative component associated with the informative amd misinformative exclusions respectively. In [39] we proposed four postulates for such a decomposition. Before stating the postulates, it is important to note that although there is a "surprising symmetry" ([41], p. 23) between the information provided by $s$ about $t$ and the information provided by $t$ about $s$, there is nothing to suggest that the components of the decomposition should be symmetric-indeed the intuition behind the decomposition only makes sense when considering the information is considered in a directed sense. As such, directed notation will be used to explicitly denote the information provided by $s$ about $t$.

Postulate 1 (Decomposition). The pointwise information provided by s about $t$ can be decomposed into two non-negative components, such that $i(s ; t)=i_{+}(s \rightarrow t)-i_{-}(s \rightarrow t)$.

Postulate 2 (Monotonicity). For all fixed $p(s, t)$ and $p\left(s^{c}, t\right)$, the function $i_{+}(s \rightarrow t)$ is a monotonically increasing, continuous function of $p\left(t^{c}, s^{c}\right)$. For all fixed $p\left(t^{c}, s\right)$ and $p\left(t^{c}, s^{c}\right)$, the function $i_{-}(s \rightarrow t)$ is a monotonically increasing continuous function of $p\left(s^{c}, t\right)$. For all fixed $p(s, t)$ and $p\left(t^{c}, s\right)$, the functions $i_{+}(s \rightarrow t)$ and $i_{-}(s \rightarrow t)$ are monotonically increasing and decreasing functions of $p\left(t^{c}, s^{c}\right)$, respectively.

Postulate 3 (Self-Information). An event cannot misinform about itself, $i_{+}(s \rightarrow s)=i(s ; s)=-\log p(s)$.
Postulate 4 (Chain Rule). The functions $i_{+}\left(s_{1,2} \rightarrow t\right)$ and $i_{-}\left(s_{1,2} \rightarrow t\right)$ satisfy a chain rule, i.e.,

$$
\begin{aligned}
i_{+}\left(s_{1,2} \rightarrow t\right) & =i_{+}\left(s_{1} \rightarrow t\right)+i_{+}\left(s_{2} \rightarrow t \mid s_{1}\right) \\
& =i_{+}\left(s_{2} \rightarrow t\right)+i_{+}\left(s_{1} \rightarrow t \mid s_{2}\right), \\
i_{-}\left(s_{1,2} \rightarrow t\right) & =i_{-}\left(s_{1} \rightarrow t\right)+i_{-}\left(s_{2} \rightarrow t \mid s_{1}\right) \\
& =i_{-}\left(s_{2} \rightarrow t\right)+i_{-}\left(s_{1} \rightarrow t \mid s_{2}\right)
\end{aligned}
$$

In Finn and Lizier [39], we proved that these postulates lead to the following forms which are unique up to the choice of the base of the logarithm in the mutual information in Postulates 1 and 3,

$$
\begin{array}{lll}
i^{+}\left(s_{1} \rightarrow t\right) & =h\left(s_{1}\right) & =-\log p\left(s_{1}\right) \\
i^{+}\left(s_{1} \rightarrow t \mid s_{2}\right) & =h\left(s_{1} \mid s_{2}\right) & =-\log p\left(s_{1} \mid s_{2}\right) \\
i^{+}\left(s_{1,2} \rightarrow t\right) & =h\left(s_{1,2}\right) & =-\log p\left(s_{1,2}\right) \\
i^{-}\left(s_{1} \rightarrow t\right) & =h\left(s_{1} \mid t\right) & =-\log p\left(s_{1} \mid t\right), \\
i^{-}\left(s_{1} \rightarrow t \mid s_{2}\right) & =h\left(s_{1} \mid t, s_{2}\right) & =-\log p\left(s_{1} \mid t, s_{2}\right), \\
i^{-}\left(s_{1,2} \rightarrow t\right) & =h\left(s_{1,2} \mid t\right) & =-\log p\left(s_{1,2} \mid t\right) . \tag{10}
\end{array}
$$

That is, the Postulates 1-4 uniquely decompose the pointwise information provided by $s$ about $t$ into the following entropic components,

$$
\begin{align*}
i(s ; t) & =i^{+}(s \rightarrow t)-i^{-}(s \rightarrow t) \\
& =h(s)-h(s \mid t) \tag{11}
\end{align*}
$$

Although the decomposition of mutual information into entropic components is well-known, it is non-trivial that Postulates 1 and 3, based on the size of the two distinct types of probability mass exclusions, lead to this particular form, but not $i(s ; t)=h(t)-h(t \mid s)$ or $i(s ; t)=h(s)+h(t)-h(s, t)$.

It is important to note that although the original motivation was to decompose the pointwise mutual information into separate components associated with informative and misinformative exclusion, the decomposition (11) does not quite possess this direct correspondence:

- The positive informational component $i^{+}(s \rightarrow t)$ does not depend on $t$ but rather only on $s$. This can be interpreted as follows: the less likely $s$ is to occur, the more specific it is when it does occur, the greater the total amount of probability mass excluded $p\left(s^{c}\right)$, and the greater the potential for $s$ to inform about $t$ (or indeed any other target realisation).
- The negative informational component $i^{-}(s \rightarrow t)$ depends on both $s$ and $t$, and can be interpreted as follows: the less likely $s$ is to coincide with the event $t$, the more uncertainty in $s$ given $t$, the greater size of the misinformative probability mass exclusion $p\left(s^{c}, t\right)$, and therefore the greater the potential for $s$ to misinform about $t$.
In other words, although the negative informational component $i^{-}(s \rightarrow t)$ does correspond directly to the size of the misinformative exclusion $p\left(s^{c}, t\right)$, the positive informational component $i^{+}(s \rightarrow t)$ does not correspond directly to the size of the informative exclusion $p\left(t^{c}, s^{c}\right)$. Rather, the positive informational component $i^{+}(s \rightarrow t)$ corresponds to the total size of the probability mass exclusions $p\left(s^{\mathrm{c}}\right)$, which is the sum of the sum of the informative and misinformative exclusions. For the sake of brevity, the positive informational component $i^{+}(s \rightarrow t)$ will be referred to as the specificity, while the negative informational component $i^{-}(s \rightarrow t)$ will be referred to as the ambiguity. The term ambiguity is due to Shannon: "[equivocation] measures the average ambiguity of the received signal" ([42], p. 67). Specificity is an antonym of ambiguity and the usage here is inline with the definition since the more specific an event $s$, the more information it could provide about $t$ after the ambiguity is taken into account.


### 3.3. Operational Interpretation of Redundant Information

Arguing about whether one piece of information differs from another piece of information is nonsensical without some kind of unambiguous definition of what it means for two pieces of information to be the same. As such, Bertschinger et al. [11] advocate the need to provide an operational interpretation of what it means for information to be unique or redundant. This section provides our operational definition of what it means for information to be the same. This definition provides a concrete interpretation of what it means for information to be redundant in terms of overlapping probability mass exclusions.

The operational interpretation of redundancy adopted here is based upon the following idea: since the pointwise information is ultimately derived from probability mass exclusions, the same information must induce the same exclusions. More formally, the information provided by a set of predictor events $s_{1}, \ldots, s_{k}$ about a target event $t$ must be the same information if each source event induces the same exclusions with respect to the two-event partition $\mathcal{T}^{t}=\left\{t, t^{c}\right\}$. While this statement makes the motivational intuition clear, it is not yet sufficient to serve as an operational interpretation of redundancy: there is no reference to the two distinct types of probability mass exclusions, the specific reference to the pointwise event space $\mathcal{T}^{t}$ has not been explained, and there is no reference to the fact the exclusions from each source may differ in size.

Informative exclusions are fundamentally different from misinformative exclusions and hence each type of exclusion should be compared separately: informative exclusions can overlap with informative exclusions, and misinformative exclusions can overlap with misinformative exclusions. In information-theoretic terms, this means comparing the specificity and the ambiguity of the sources separately-i.e., considering a measure of redundant specificity and a separate measure of redundant ambiguity. Crucially, these quantities (being pointwise entropies) are unsigned meaning that the difficulties associated with Axiom 2 (Monotonicity) and signed pointwise mutual information in Section 2.3 will not be an issue here.

The specific reference to the two-event partition $\mathcal{T}^{t}$ in the above statement is based upon Remark 1 and is crucially important. The pointwise mutual information does not depend on the apportionment of
the informative exclusions across the set of events which did not occur, hence the pointwise redundant information should not depend on this apportionment either. In other words, it is immaterial if two predictor events $s_{1}$ and $s_{2}$ exclude different elementary events within the target complementary event $t^{c}$ (assuming the probability mass excluded is equal) since with respect to the realised target event $t$ the difference between the exclusions is only semantic. This has important implications for the comparison of exclusions from different predictor events. As the pointwise mutual information depends on, and only depends on, the size of the exclusions, then the only sensible comparison is a comparison of size. Hence, the common or overlapping exclusion must be the smallest exclusion. Thus, consider the following operational interpretation of redundancy:

Operational Interpretation (Redundant Specificity). The redundant specificity between a set of predictor events $s_{1}, \ldots, s_{n}$ is the specificity associated with the source event which induces the smallest total exclusions.

Operational Interpretation (Redundant Ambiguity). The redundant ambiguity between a set of predictor events $s_{1}, \ldots, s_{n}$ is the ambiguity associated with the source event which induces the smallest misinformative exclusion.

### 3.4. Motivational Example

To motivate the above operational interpretation, and in particular the need to treat the specificity separately to the ambiguity, consider Figure 2. In this pointwise example, two different predictor events provide the same amount of pointwise information since $P\left(T \mid s_{1}^{1}\right)=P\left(T \mid s_{2}^{1}\right)$, and yet the information provided by each event is in some way different since each excludes different sections of the target distribution $P(T)$. In particular, $s_{1}^{1}$ and $s_{2}^{1}$ both preclude the target event $t^{2}$, while $s_{2}^{1}$ additionally excludes probability mass associated with target events $t^{1}$ and $t^{3}$. From the perspective of the pointwise mutual information the events $s_{1}^{1}$ and $s_{2}^{1}$ seem to be providing the same information as

$$
\begin{equation*}
i\left(s_{1}^{1} \rightarrow t^{1}\right)=i\left(s_{2}^{1} \rightarrow t^{1}\right)=\log 4 / 3 \text { bit. } \tag{12}
\end{equation*}
$$

However, from the perspective of the specificity and the ambiguity it can be seen that information is being provided in different ways since

$$
\begin{align*}
i^{+}\left(s_{1}^{1} \rightarrow t^{1}\right) & =\log 4 / 3 \mathrm{bit},
\end{align*} i^{-}\left(s_{1}^{1} \rightarrow t^{1}\right)=0 \mathrm{bit}, ~ 子, ~=1 \mathrm{bit.} .
$$

Now consider the problem of decomposing information into its unique, redundant and complementary components. Figure 2 shows where exclusions induced by $s_{1}^{1}$ and $s_{2}^{1}$ overlap where they both exclude the target event $t^{2}$ which is an informative exclusion. This is the only exclusion induced by $s_{1}^{1}$ and hence all of the information associated with this exclusion must be redundantly provided by the event $s_{2}^{1}$. Without any formal framework, consider taking the redundant specificity and redundant ambiguity,

$$
\begin{align*}
& r^{+}\left(s_{1}^{1}, s_{2}^{1} \rightarrow t^{1}\right)=i^{+}\left(s_{1}^{1} \rightarrow t^{1}\right)=\log 4 / 3 \mathrm{bit}  \tag{14}\\
& r^{-}\left(s_{1}^{1}, s_{2}^{1} \rightarrow t_{1}\right)=i^{-}\left(s_{1}^{1} \rightarrow t^{1}\right)=0 \mathrm{bit} . \tag{15}
\end{align*}
$$

This would mean that the event $s_{2}^{1}$ provides the following unique specificity and unique ambiguity,

$$
\begin{align*}
& u^{+}\left(s_{1}^{1} \backslash s_{2}^{1} \rightarrow t^{1}\right)=i^{+}\left(s_{1}^{1} \rightarrow t^{1}\right)-r^{+}\left(s_{1}^{1}, s_{2}^{1} \rightarrow t^{1}\right)=1 \mathrm{bit},  \tag{16}\\
& u^{-}\left(s_{1}^{1} \backslash s_{2}^{1} \rightarrow t^{1}\right)=i^{-}\left(s_{1}^{1} \rightarrow t^{1}\right)-r^{-}\left(s_{1}^{1}, s_{2}^{1} \rightarrow t^{1}\right)=1 \mathrm{bit} . \tag{17}
\end{align*}
$$

The redundant specificity $\log 4 / 3$ bit accounts for the overlapping informative exclusion of the event $t^{2}$. The unique specificity and unique ambiguity from $s_{2}^{1}$ are associated with its non-overlapping informative and misinformative exclusions; however, both of these 1 bit and hence, on net, $s_{2}^{1}$ is no
more informative than $s_{1}^{1}$. Although obtained without a formal framework, this example highlights a need to consider the specificity and ambiguity rather than merely the pointwise mutual information.


Figure 2. Sample probability mass diagrams for two predictors $S_{1}$ and $S_{2}$ to a given target $T$. Here events in the two different predictor spaces provide the same amount of pointwise information about the target event, $\log _{2} 4 / 3$ bits, since $P\left(T \mid s_{1}^{1}\right)=P\left(T \mid s_{2}^{1}\right)$, although each excludes different sections of the target distribution $P(T)$. Since they both provide the same amount of information, is there a way to characterise what information the additional unique exclusions from the event $s_{2}^{1}$ are providing?

## 4. Pointwise Partial Information Decomposition Using Specificity and Ambiguity

Based upon the argumentation of Section 3, consider the following axioms:
Axiom 1 (Symmetry). Pointwise redundant specificity $i_{\cap}^{+}$and pointwise redundant ambiguity $i_{\cap}^{-}$are invariant under any permutation $\sigma$ of source events,

$$
\begin{aligned}
& i_{\cap}^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t\right)=i_{\cap}^{+}\left(\sigma\left(\boldsymbol{a}_{1}\right), \ldots, \sigma\left(\boldsymbol{a}_{k}\right) \rightarrow t\right) \\
& i_{\cap}^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t\right)=i_{\cap}^{-}\left(\sigma\left(\boldsymbol{a}_{1}\right), \ldots, \sigma\left(\boldsymbol{a}_{k}\right) \rightarrow t\right)
\end{aligned}
$$

Axiom 2 (Monotonicity). Pointwise redundant specificity $i_{\cap}^{+}$and pointwise redundant ambiguity $i_{\cap}^{-}$decreases monotonically as more source events are included,

$$
\begin{aligned}
& i_{\cap}^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k-1}, \boldsymbol{a}_{k} \rightarrow t\right) \leq i_{\cap}^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k-1} \rightarrow t\right) \\
& i_{\cap}^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k-1}, \boldsymbol{a}_{k} \rightarrow t\right) \leq i_{\cap}^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k-1} \rightarrow t\right)
\end{aligned}
$$

with equality if $\boldsymbol{a}_{k} \supseteq \boldsymbol{a}_{i}$ for any $\boldsymbol{a}_{i} \in\left\{\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k-1}\right\}$.
Axiom 3 (Self-redundancy). Pointwise redundant specificity $i_{\cap}^{+}$and pointwise redundant ambiguity $i_{\cap}^{-}$for a single source event $\boldsymbol{a}_{i}$ equals the specificity and ambiguity respectively,

$$
\begin{aligned}
& i_{\cap}^{+}\left(\boldsymbol{a}_{i} \rightarrow t\right)=i^{+}\left(\boldsymbol{a}_{i} \rightarrow t\right)=h\left(\boldsymbol{a}_{i}\right) \\
& i_{\cap}^{-}\left(\boldsymbol{a}_{i} \rightarrow t\right)=i^{-}\left(\boldsymbol{a}_{i} \rightarrow t\right)=h\left(\boldsymbol{a}_{i} \mid t\right)
\end{aligned}
$$

As shown in Appendix B.1, Axioms 1-3 induce two lattices-namely the specificity lattice and ambiguity lattice-which are depicted in Figure 3. Furthermore, each lattice is defined for every discrete realisation from $P\left(S_{1}, \ldots, S_{n}, T\right)$. The redundancy measures $i_{\cap}^{+}$or $i_{\cap}^{-}$can be thought of as a cumulative information functions which integrate the specificity or ambiguity uniquely contributed by each node as one moves up each lattice. Finally, just as in PID, performing a Möbius inversion over each lattice yielding the unique contributions of specificity and ambiguity from each sources event.

Similarly to PID, the specificity and ambiguity lattices provide a structure for information decomposition, but unique evaluation requires a separate definition of redundancy. However, unlike PID (or even PPID), this evaluation requires both a definition of pointwise redundant specificity and pointwise redundant ambiguity. Before providing these definitions, it is helpful to first see how the specificity and ambiguity lattices can be used to decompose multivariate information in the now familiar bivariate case.


Figure 3. The lattice induced by the partial order $\preceq$ (A15) over the sources $\mathscr{A}(\boldsymbol{s})$ (A14). (Left) the lattice for $s=\left\{s_{1}, s_{2}\right\}$; (Right) the lattice for $s=\left\{s_{1}, s_{2}, s_{3}\right\}$. See Appendix B for further details. Each node corresponds to the self-redundancy (Axiom 3) of a source event, e.g., $\{1\}$ corresponds to the source event $\left\{\left\{s_{1}\right\}\right\}$, while $\{12,13\}$ corresponds to the source event $\left\{\left\{s_{1,2}\right\},\left\{s_{1,3}\right\}\right\}$. Note that the specificity and ambiguity lattices share the same structure as the redundancy lattice of partial information decomposition (PID) (cf. Figure 2 in [1]).

### 4.1. Bivariate PPID Using the Specificity and Ambiguity

Consider again the bivariate case where the aim is to decompose the information provided by $s_{1}$ and $s_{2}$ about $t$. The specificity lattice can be used to decompose the pointwise specificity,

$$
\begin{align*}
i^{+}\left(s_{1,2} \rightarrow t\right) & =r^{+}\left(s_{1}, s_{2} \rightarrow t\right)+u^{+}\left(s_{1} \backslash s_{2} \rightarrow t\right)+u^{+}\left(s_{2} \backslash s_{1} \rightarrow t\right)+c^{+}\left(s_{1}, s_{2} \rightarrow t\right), \\
i^{+}\left(s_{1} \rightarrow t\right) & =r^{+}\left(s_{1}, s_{2} \rightarrow t\right)+u^{+}\left(s_{1} \backslash s_{2} \rightarrow t\right),  \tag{18}\\
i^{+}\left(s_{2} \rightarrow t\right) & =r^{+}\left(s_{1}, s_{2} \rightarrow t\right)+u^{+}\left(s_{2} \backslash s_{1} \rightarrow t\right) ;
\end{align*}
$$

while the ambiguity lattice can be used to decompose the pointwise ambiguity,

$$
\begin{align*}
i^{-}\left(s_{1,2} \rightarrow t\right) & =r^{-}\left(s_{1}, s_{2} \rightarrow t\right)+u^{-}\left(s_{1} \backslash s_{2} \rightarrow t\right)+u^{-}\left(s_{2} \backslash s_{1} \rightarrow t\right)+c^{-}\left(s_{1}, s_{2} \rightarrow t\right), \\
i^{-}\left(s_{1} \rightarrow t\right) & =r^{-}\left(s_{1}, s_{2} \rightarrow t\right)+u^{-}\left(s_{1} \backslash s_{2} \rightarrow t\right),  \tag{19}\\
i^{-}\left(s_{2} \rightarrow t\right) & =r^{-}\left(s_{1}, s_{2} \rightarrow t\right)+u^{-}\left(s_{2} \backslash s_{1} \rightarrow t\right) .
\end{align*}
$$

These equations share the same structural form as (3) only now decompose the specificity and the ambiguity rather than the pointwise mutual information, e.g., $r^{+}\left(s_{1}, s_{2} \rightarrow t\right)$ denotes the redundant specificity while $u^{-}\left(s_{1} \backslash s_{2} \rightarrow t\right)$ denoted the unique ambiguity from $s_{1}$. Just as in for (3), this decomposition could be considered for every discrete realisation on the support of the joint distribution $P\left(S_{1}, S_{2}, T\right)$.

There are two ways one can be combine these values. Firstly, in a similar manner to (4), one could take the expectation of the atoms of specificity, or the atoms of ambiguity, over all discrete realisations yielding the average PI atoms of specificity and ambiguity,

$$
\begin{array}{rlll}
U^{+}\left(S_{1} \backslash S_{2} \rightarrow T\right) & =\left\langle u^{+}\left(s_{1} \backslash s_{2} \rightarrow t\right)\right\rangle, & U^{-}\left(S_{1} \backslash S_{2} \rightarrow T\right) & =\left\langle u^{-}\left(s_{1} \backslash s_{2} \rightarrow t\right)\right\rangle, \\
U^{+}\left(S_{2} \backslash S_{1} \rightarrow T\right) & =\left\langle u^{+}\left(s_{2} \backslash s_{1} \rightarrow t\right)\right\rangle, & U^{-}\left(S_{2} \backslash S_{1} \rightarrow T\right) & =\left\langle u^{-}\left(s_{2} \backslash s_{1} \rightarrow t\right)\right\rangle, \\
R^{+}\left(S_{1}, S_{2} \rightarrow T\right) & =\left\langle r^{+}\left(s_{1}, s_{2} \rightarrow t\right)\right\rangle, & R^{-}\left(S_{1}, S_{2} \rightarrow T\right) & =\left\langle r^{-}\left(s_{1}, s_{2} \rightarrow t\right)\right\rangle,  \tag{20}\\
C^{+}\left(S_{1}, S_{2} \rightarrow T\right) & =\left\langle c^{+}\left(s_{1}, s_{2} \rightarrow t\right)\right\rangle . & C^{-}\left(s_{1}, S_{2} \rightarrow T\right) & =\left\langle c^{-}\left(s_{1}, s_{2} \rightarrow t\right)\right\rangle .
\end{array}
$$

Alternatively, one could subtract the pointwise unique, redundant and complementary ambiguity from the pointwise unique, redundant and complementary specificity yielding the pointwise unique, pointwise redundant and pointwise complementary information, i.e., recover the atoms from PPID,

$$
\begin{align*}
r\left(s_{1}, s_{2} \rightarrow t\right) & =r^{+}\left(s_{1}, s_{2} \rightarrow t\right)-r^{-}\left(s_{1}, s_{2} \rightarrow t\right), \\
u\left(s_{1} \backslash s_{2} \rightarrow t\right) & =u^{+}\left(s_{1} \backslash s_{2} \rightarrow t\right)-u^{-}\left(s_{1} \backslash s_{2} \rightarrow t\right), \\
u\left(s_{2} \backslash s_{1} \rightarrow t\right) & =u^{+}\left(s_{2} \backslash s_{1} \rightarrow t\right)-u^{-}\left(s_{2} \backslash s_{1} \rightarrow t\right),  \tag{21}\\
c\left(s_{1}, s_{2} \rightarrow t\right) & =c^{+}\left(s_{1}, s_{2} \rightarrow t\right)-c^{-}\left(s_{1}, s_{2} \rightarrow t\right) .
\end{align*}
$$

Both (20) and (21) are linear operations, hence one could perform both of these operations (in either order) to obtain the average unique, average redundant and average complementary information, i.e., recover the atoms from PID,

$$
\begin{align*}
R\left(S_{1}, S_{2} \rightarrow T\right) & =R^{+}\left(S_{1}, S_{2} \rightarrow T\right)-R^{-}\left(S_{1}, S_{2} \rightarrow T\right), \\
U\left(S_{1} \backslash S_{2} \rightarrow T\right) & =U^{+}\left(S_{1} \backslash S_{2} \rightarrow T\right)-U^{-}\left(S_{1} \backslash S_{2} \rightarrow T\right), \\
U\left(S_{2} \backslash S_{1} \rightarrow T\right) & =U^{+}\left(S_{2} \backslash S_{1} \rightarrow T\right)-U^{-}\left(S_{2} \backslash S_{1} \rightarrow T\right),  \tag{22}\\
C\left(S_{1}, S_{2} \rightarrow T\right) & =C^{+}\left(S_{1}, S_{2} \rightarrow T\right)-C^{-}\left(S_{1}, S_{2} \rightarrow T\right) .
\end{align*}
$$

### 4.2. Redundancy Measures on the Specificity and Ambiguity Lattices

Now that we have a structure for our information decomposition, there is a need to provide a definition of the pointwise redundant specificity and pointwise redundant ambiguity. However, before attempting to provide such a definition, there is a need to consider Remark 1 and the operational interpretation of in Section 3.3. In particular, the pointwise redundant specificity $i_{\cap}^{+}$and pointwise redundant ambiguity $i_{\cap}^{-}$should only depend on the size of informative and misinformative exclusions. They should not depend on the apportionment of the informative exclusions across the set of elementary events contained in the complementary event $t^{c}$. Formally, this requirement will be enshrined via the following axiom.

Axiom 4 (Two-event Partition). The pointwise redundant specificity $i_{\cap}^{+}$and pointwise redundant ambiguity $i_{\cap}^{-}$are functions of the probability measures on the two-event partitions $\mathcal{A}_{1}^{a_{1}} \times \mathcal{T}^{t}, \ldots, \mathcal{A}_{k}^{a_{k}} \times \mathcal{T}^{t}$.

Since the pointwise redundant specificity $i_{\cap}^{+}$is specificity associated with the source event which induces the smallest total exclusions, and pointwise redundant ambiguity $i_{\cap}^{-}$is the ambiguity associated with the source event which induces the smallest misinformative exclusion, consider the following definitions.

Definition 1. The pointwise redundant specificity is given by

$$
\begin{equation*}
r_{\min }^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t\right)=\min _{\boldsymbol{a}_{i}} i^{+}\left(\boldsymbol{a}_{i} \rightarrow t\right)=\min _{\boldsymbol{a}_{i}} h\left(\boldsymbol{a}_{i}\right) . \tag{23}
\end{equation*}
$$

Definition 2. The pointwise redundant ambiguity is given by

$$
\begin{equation*}
r_{\min }^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t\right)=\min _{\boldsymbol{a}_{i}} i^{-}\left(\boldsymbol{a}_{i} \rightarrow t\right)=\min _{\boldsymbol{a}_{j}} h\left(\boldsymbol{a}_{j} \mid t\right) \tag{24}
\end{equation*}
$$

Theorem 1. The definitions of $r_{\text {min }}^{+}$and $r_{\text {min }}^{-}$satisfy Axioms 1-4.
Theorem 2. The redundancy measures $r_{\min }^{+}$and $r_{\min }^{-}$increase monotonically on the $\langle\mathscr{A}(\boldsymbol{s}), \preceq\rangle$.
Theorem 3. The atoms of partial specificity $\pi^{+}$and partial ambiguity $\pi^{-}$evaluated using the measures $r_{\text {min }}^{+}$ and $r_{\text {min }}^{-}$on the specificity and ambiguity lattices (respectively), are non-negative.

Appendix B. 2 contains the proof of Theorems 1-3 and further relevant consideration of Defintions 1 and 2. As in (20), one can take the expectation of the either the pointwise redundant specificity $r_{\text {min }}^{+}$or the pointwise redundant ambiguity $r_{\text {min }}^{-}$to get the average redundant specificity $R_{\text {min }}^{+}$or the average redundant ambiguity $R_{\min }^{-}$. Alternatively, just as in (21), one can recombine the pointwise redundant specificity $r_{\min }^{+}$and the pointwise redundant ambiguity $r_{\min }^{-}$to get the pointwise redundant information $r_{\text {min }}$. Finally, as per (22), one could perform both of these (linear) operations in either order to obtain the average redundant information $R_{\min }$. Note that while Theorem 3 proves that the atoms of partial specificity $\pi^{+}$and partial ambiguity $\pi^{-}$are non-negative, it is trivial to see that $r_{\min }$ could be negative since when source events can redundantly provide misinformation about a target event. As shown in the following theorem, $R_{\min }$ can also be negative.

Theorem 4. The atoms of partial average information $\Pi$ evaluated by recombining and averaging $\pi^{ \pm}$are not non-negative.

This means that the measure $R_{\min }$ does not satisfy local positivity. Nonetheless the negativity of $R_{\text {min }}$ is readily explainable in terms of the operational interpretation of Section 3.3, as will be discussed further in Section 5.4. However, failing to satisfy local positivity does mean that $r_{\text {min }}$ and $R_{\min }$ do not satisfy the target monotonicity property first discussed in Bertschinger et al. [5]. Despite this, as the following theorem shows, the measures do satisfy the target chain rule.

Theorem 5 (Pointwise Target Chain Rule). Given the joint target realisation $t_{1,2}$, the pointwise redundant information $r_{\text {min }}$ satisfies the following chain rule,

$$
\begin{align*}
r_{\text {min }}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1,2}\right) & =r_{\text {min }}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1}\right)+r_{\text {min }}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{2} \mid t_{1}\right), \\
& =r_{\text {min }}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{2}\right)+r_{\text {min }}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1} \mid t_{2}\right) . \tag{25}
\end{align*}
$$

The proof of the last theorem is deferred to Appendix B.3. Note that since the expectation is a linear operation, Theorem 5 also holds for the average redundant information $R_{\min }$. Furthermore, as these results apply to any of the source events, the target chain rule will hold for any of the PPI atoms, e.g., (21), and any of the PI atoms, e.g., (22). However, no such rule holds for the pointwise redundant specificity or ambiguity. The specificity depends only on the predictor event, i.e., does not depend on the target events. As such, when an increasing number of target events are considered, the specificity remains unchanged. Hence, a target chain rule cannot hold for the specificity, or the ambiguity alone.

## 5. Discussion

PPID using the specificity and ambiguity takes the ideas underpinning PID and applies them on a pointwise scale while circumventing the monotonicity issue associated with the signed pointwise mutual information. This section will explore the various properties of the decomposition in an example driven manner and compare the results to the most widely-used measures from the existing PID literature. (Further examples can be found in Appendix C.) The following shorthand notation will be utilised in the figures throughout this section:

$$
\begin{array}{lll}
i_{1}^{+}=i^{+}\left(s_{1} \rightarrow t\right), & i_{2}^{+}=i^{+}\left(s_{2} \rightarrow t\right), & i_{1,2}^{+}=i^{+}\left(s_{1,2} \rightarrow t\right), \\
i_{1}^{-}=i^{-}\left(s_{1} \rightarrow t\right), & i_{2}^{-}=i^{-}\left(s_{2} \rightarrow t\right), & i_{1,2}^{-}=i^{-}\left(s_{1,2} \rightarrow t\right), \\
u_{1}^{+}=u^{+}\left(s_{1} \backslash s_{2} \rightarrow t\right), & u_{2}^{+}=u^{+}\left(s_{2} \backslash s_{1} \rightarrow t\right), & r^{+}=r^{+}\left(s_{1}, s_{2} \rightarrow t\right), \\
u_{1}^{-}=u^{-}\left(s_{1} \backslash s_{2} \rightarrow t\right), & u_{2}^{-}=u^{+}\left(s_{1}, s_{2} \rightarrow t\right), \\
\left.s_{2} \backslash s_{1} \rightarrow t\right), & r^{-}=r^{-}\left(s_{1}, s_{2} \rightarrow t\right), & c^{-}=c^{-}\left(s_{1}, s_{2} \rightarrow t\right) .
\end{array}
$$

### 5.1. Comparison to Existing Measures

A similar approach to the decomposition presented in this paper is due to Ince [18], who also sought to define a pointwise information decomposition. Despite the similarity in this regard, the redundancy measure $I_{\mathrm{CCS}}$ presented in [18] approaches the pointwise monotonicity problem of Section 2.3 in a different way to the decomposition presented in this paper. Specifically, $I_{\mathrm{CCS}}$ aims to utilise the pointwise co-information as a measure of pointwise redundant information since it "quantifies the set-theoretic overlap of the two univariate [pointwise] information values" ([18], p. 14). There are, however, difficulties with this approach. Firstly (unlike the average mutual information and the Shannon inequalities), there are no inequalities which support this interpretation of pointwise co-information as the set-theoretic overlap of the univariate pointwise information terms-indeed, both the univariate pointwise information and the pointwise co-information are signed measures. Secondly, the pointwise co-information conflates the pointwise redundant information with the pointwise complementary information, since by (3) we have that

$$
\begin{equation*}
\operatorname{co-i}\left(s_{1} ; s_{2} ; t\right):=i\left(s_{1} ; t\right)+i\left(s_{2} ; t\right)-i\left(s_{1,2}, t\right)=r\left(s_{1}, s_{2} \rightarrow t\right)-c\left(s_{1}, s_{2} \rightarrow t\right) \tag{26}
\end{equation*}
$$

Aware of these difficulties, Ince defines $I_{C C S}$ such that it only interprets the pointwise co-information as a measure of set-theoretic overlap in the case where all three pointwise information terms have the same sign, arguing that these are the only situations which admit a clear interpretation in terms of a common change in surprisal. In the other difficult to interpret situations, $I_{\mathrm{CCS}}$ defines the pointwise redundant information to be zero. This approach effectively assumes that $c\left(s_{1}, s_{2} \rightarrow t\right)=0$ in (26) when $i\left(s_{1} ; t\right), i\left(s_{2} ; t\right)$ and $c o-i\left(s_{1} ; s_{2} ; t\right)$ all have the same sign.

In a subsequent paper, Ince [19] also presented a partial entropy decomposition which aims to decompose multivariate entropy rather than multivariate information. As such, this decomposition is more similar to PPID using specificity and ambiguity than Ince's aforementioned decomposition. Although similar in this regard, the measure of pointwise redundant entropy $H_{\text {cs }}$ presented in [19] takes a different approach to the one presented in this paper. Specifically, $H_{\mathrm{cs}}$ also uses the pointwise co-information as a measure of set-theoretic overlap and hence as a measure of pointwise redundant entropy. As the pointwise entropy is unsigned, the difficulties are reduced but remain present due to the signed pointwise co-information. In a manner similar to $I_{\mathrm{CCS}}$, Ince defines $H_{\mathrm{CS}}$ such that it only interprets the pointwise co-information as a measure of set-theoretic overlap when it is positive. As per $I_{\mathrm{CCS}}$, this effectively assumes that $c\left(s_{1}, s_{2} \rightarrow t\right)=0$ in (26) when all information terms have the same sign. When the pointwise co-information is negative, $H_{\mathrm{cs}}$ simply ignores the co-information by defining the pointwise redundant information to be zero. In contrast to both of Ince's approaches, PPID using specificity and ambiguity does not dispose of the set-theoretic intuition in these difficult to interpret situations. Rather, our approach considers the notion of redundancy in terms of overlapping exclusions-i.e., in terms of the underlying, unsigned measures which are amenable to a set-theoretic interpretation.

The measures of pointwise redundant specificity $r_{\min }^{+}$and pointwise redundant ambiguity $r_{\min }^{-}$ from Definitions 1 and 2 are also similar to both the minimum mutual information $I_{\text {MMI }}$ [17] and the original PID redundancy measure $I_{\min }$ [1]. Specifically, all three of these approaches consider the redundant information to be the minimum information provided about a target event $t$. The difference is that $I_{\min }$ applies this idea to the sources $A_{1}, \ldots, A_{k}$, i.e., to collections of entire predictor variables from $S$, while $r_{\min }^{ \pm}$apply this notion to the source events $a_{1}, \ldots, a_{k}$, i.e., to collections of predictor events from $s$. In other words, while the measure $I_{\min }$ can be regarded as being semi-pointwise (since it considers the information provided by the variables $S_{1}, \ldots, S_{n}$ about an event $t$ ), the measures $r_{\text {min }}^{ \pm}$ are fully pointwise (since they consider the information provided by the events $s_{1}, \ldots, s_{n}$ about an event $t$ ). This difference in approach is most apparent in the probability distribution PWUNQ-unlike PID, PPID using the specificity and ambiguity respects the pointwise nature of information, as we will see in Section 5.3.

PPID using specificity and ambiguity also share certain similarities with the bivariate PID induced by the measure $\widetilde{U I}$ of Bertschinger et al. [11]. Firstly, Axiom 4 can be considered to be a pointwise adaptation of their Assumption (*), i.e., the measures $r_{\min }^{ \pm}$depend only on the marginal distributions $P\left(S_{1}, T\right)$ and $P\left(S_{2}, T\right)$ with respect to the two-event partitions $\mathcal{S}_{1}^{s_{1}} \times \mathcal{T}^{t}$ and $\mathcal{S}_{2}^{s_{2}} \times \mathcal{T}^{t}$. Secondly, in PPID using specificity and ambiguity, the only way one can only decide if there is complementary information $c\left(s_{1}, s_{2} \rightarrow t\right)$ is by knowing the joint distribution $P\left(S_{1}, S_{2}, T\right)$ with respect to the joint two-event partitions $\mathcal{S}_{1}^{s_{1}} \times \mathcal{S}_{2}^{s_{2}} \times \mathcal{T}^{t}$. This is (in effect) a pointwise form of their Assumption ( $* *$ ). Thirdly, by definition $r_{\min }^{ \pm}$are given by the minimum value that any one source event provides. This is the largest possible value that one could take for these quantities whilst still requiring that the unique specificity and ambiguity be non-negative. Hence, within each discrete realisation, $r_{\min }^{ \pm}$minimise the unique specificity and ambiguity whilst maximising the redundant specificity and ambiguity. This is similar to $\widetilde{U I}$ which minimises the (average) unique information while still satisfying Assumption $(*)$. Finally, note that since the measure $\mathcal{S}_{\mathrm{VK}}$ produces a bivariate decomposition which is equivalent to that of $\widetilde{U I}[11]$, the same similarities apply between PPID using specificity and ambiguity and the decomposition induced by $\mathcal{S}_{\mathrm{VK}}$ from Griffith and Koch [12].

### 5.2. Probability Distribution XOR

Figure 4 shows the canonical example of synergy, exclusive-or (XOR) which considers two independently distributed binary predictor variables $S_{1}$ and $S_{2}$ and a target variable $T=S_{1}$ XOR $S_{2}$. There are several important points to note about the decomposition of XOR. Firstly, despite providing zero pointwise information, an individual predictor event does indeed induce exclusions. However, the informative and misinformative exclusions are perfectly balanced such that the posterior (conditional) distribution is equal to the prior distribution, e.g., see the red coloured exclusions induced by $S_{1}=0$ in Figure 4. In information-theoretic terms, for each realisation, the pointwise specificity equals 1 bit since half of the total probability mass remains while the pointwise ambiguity also equals 1 bit since half of the probability mass associated with the event which subsequently occurs (i.e., $T=0$ ), remains. These are perfectly balanced such that when recombined, as per (11), the pointwise mutual information is equal to 0 bit , as one would expect.

Secondly, $S_{1}=0$ and $S_{2}=0$ both induce the same exclusions with respect to the target pointwise event space $\mathcal{T}^{T=0}$. Hence, as per the operational interpretation of redundancy adopted in Section 3.3, there is 1 bit of pointwise redundant specificity and 1 bit of pointwise redundant ambiguity in each realisation. The presence of (a form of) redundancy in XOR is novel amongst the existing measures in the PID literature. (Ince [19] also identifies a form of redundancy in XOR.) Thirdly, despite the presence of this redundancy, recombining the atoms of pointwise specificity and ambiguity for each realisation, as per (21), leaves only one non-zero PPI atom: namely the pointwise complementary information $c\left(s_{1}, s_{2} \rightarrow t\right)=1$ bit. Furthermore, this is true for every pointwise realisation and hence, by (22), the only non-zero PI atom is the average complementary information $C\left(S_{1}, S_{2} \rightarrow T\right)=1$ bit.


| $p$ | $s_{1}$ | $s_{2}$ | $t$ | $i_{1}^{+}$ | $i_{1}^{-}$ | $i_{2}^{+}$ | $i_{2}^{-}$ | $i_{12}^{+}$ | $i_{12}^{-}$ | $r^{+}$ | $u_{1}^{+}$ | $u_{2}^{+}$ | $c^{+}$ | $r$ | $u_{1}$ | $u_{2}$ | $c^{-}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1/4 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 1/4 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 1/4 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 1/4 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| Expected values |  |  |  | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |

$R\left(S_{1}, S_{2} \rightarrow T\right)=0$ bit $\quad U\left(S_{1} \backslash S_{2} \rightarrow T\right)=0$ bit $\quad U\left(S_{2} \backslash S_{1} \rightarrow T\right)=0$ bit $\quad C\left(S_{1}, S_{2} \rightarrow T\right)=1$ bit
Figure 4. Example XOR. (Top) probability mass diagrams for the realisation ( $S_{1}=0, S_{2}=0, T=0$ ); (Middle) For each realisation, the pointwise specificity and pointwise ambiguity has been evaluated using (5) and (8) respectively. The pointwise redundant specificity and pointwise redundant ambiguity are then determined using (23) and (24). The decomposition is calculated using (18) and (19). The expected specificity and ambiguity are calculated with (20); (Bottom) The average information is given by (22). As expected, XOR yields 1 bit of complementary information.

### 5.3. Probability Distribution PwUNQ

Figure 5 shows the probability distribution PWUNQ introduced in Section 2.2. Recombining the decomposition via (21) yields the pointwise information decomposition proposed in Table 1—unsurprisingly, the explicitly pointwise approach results in a decomposition which does not suffer from the pointwise unique problem of Section 2.2.


| $p$ | $s_{1}$ | $s_{2}$ | $t$ | $i_{1}^{+}$ | $i_{1}^{-}$ | $i_{2}^{+}$ | $i_{2}^{-}$ | $i_{12}^{+}$ | $i_{12}^{-}$ | $r^{+}$ | $u_{1}^{+}$ | $u_{2}^{+}$ | $c^{+}$ | $r^{-}$ | $u_{1}$ | $u_{2}^{-}$ | $c^{-}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1/4 | 0 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 1/4 | 1 | 0 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 1/4 | 0 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 1/4 | 2 | 0 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| Expected values |  |  |  | 3/2 | 1 | 3/2 | 1 | 2 | 1 | 1 | 1/2 | 1/2 | 0 | 1 | 0 | 0 | 0 |

$R\left(S_{1}, S_{2} \rightarrow T\right)=0$ bit $\quad U\left(S_{1} \backslash S_{2} \rightarrow T\right)=1 / 2$ bit $\quad U\left(S_{2} \backslash S_{1} \rightarrow T\right)=1 / 2$ bit $\quad C\left(S_{1}, S_{2} \rightarrow T\right)=0$ bit
Figure 5. Example PWUNQ. (Top) probability mass diagrams for the realisation ( $S_{1}=0, S_{2}=1, T=1$ ); (Middle) For each realisation, the pointwise partial information decomposition (PPID) using specificity and ambiguity is evaluated (see Figure 4 for details). Upon recombination as per (21), the PPI decomposition from Table 1 is attained; (Bottom) as does the average information-the decomposition does not have the pointwise unique problem.

In each realisation, observing a 0 in either source provides the same balanced informative and misinformative exclusions as in XOR. Observing either a 1 or 2 provides the same misinformative exclusion as observing the 0 , but provides a larger informative exclusion than 0 . This leaves only the probability mass associated with the event which subsequently occurs remaining (hence why observing a 1 and 2 is fully informative about the target). Information theoretically, in each realisation the predictor events provide 1 bit of redundant pointwise specificity and 1 bit of redundant pointwise ambiguity while the fully informative event additionally provides 1 bit of unique specificity.

### 5.4. Probability Distribution RDNERR

Figure 6 shows the probability distribution redundant-error (RDNERR) which considers two predictors which are nominally redundant and fully informative about the target, but where one predictor occasionally makes an erroneous prediction. Specifically, Figure 6 shows the decomposition of RDNERR where $S_{2}$ makes an error with a probability $\varepsilon=1 / 4$. The important feature to note about this probability distribution is that upon recombining the specificity and ambiguity and taking the expectation over every realisation, the resultant average unique information from $S_{2}$ is $U\left(S_{2} \backslash S_{1} \rightarrow T\right)=-0.811$ bit.

On first inspection, the result that the average unique information can be negative may seem problematic; however, it is readily explainable in terms of the operational interpretation of Section 3.3. In RDNERR, a source event always excludes exactly $1 / 2$ of the total probability mass, thus every realisation contains 1 bit of redundant pointwise specificity. The events of the error-free $S_{1}$ induce only informative exclusions and as such provide 0 bit of pointwise ambiguity in each realisation. In contrast, the events in the error-prone $S_{2}$ always induce a misinformative exclusion, meaning that $S_{2}$ provides unique pointwise ambiguity in every realisation. Since $S_{2}$ never provides unique specificity, the average unique information is negative on average.

Despite the negativity of the average unique information, in is important to observe that $S_{2}$ provides 0.189 bit of information since $S_{2}$ also provides 1 bit of average redundant information. It is not that $S_{2}$ provides negative information on average (as this is not possible); rather it is that not all of the information provided by $S_{2}$ (i.e., the specificity) is "useful" ([42], p. 21). This is in contrast to $S_{1}$ which only provides useful specificity. To summarise, it is the unique ambiguity which distinguishes the information provided by variable $S_{2}$ from $S_{1}$, and hence why $S_{2}$ is deemed to provide negative average unique information. This form of uniqueness can only be distinguished by allowing the average unique information to be negative. This of course, requires abandoning the local positivity as a required property, as per Theorem 4. Few of the existing measures in the PID literature consider dropping this requirement as negative information quantities are typically regarded as being "unfortunate" ([43], p. 49). However, in the context of the pointwise mutual information, negative information values are readily interpretable as being misinformative values. Despite this, the average information from each predictor must be non-negative; however, it may be that what distinguishes one predictor from another are precisely the misinformative predictor events, meaning that the unique information is in actual fact, unique misinformation. Forgoing local positivity makes the PPID using specificity and ambiguity novel (the other exception in this regard is Ince [18] who was first to consider allowing negative average unique information.)


Figure 6. Example RDNERr. (Top) probability mass diagrams for the realisations ( $S_{1}=0, S_{2}=0, T=0$ ) and ( $S_{1}=0, S_{2}=1, T=0$ ); (Middle) for each realisation, the PPID using specificity and ambiguity is evaluated (see Figure 4 for details); (Bottom) the average PI atoms may be negative as the decomposition does not satisfy local positivity.

### 5.5. Probability Distribution TBC

Figure 7 shows the probability distribution two-bit-copy (TBC) which considers two independently distributed binary predictor variables $S_{1}$ and $S_{2}$, and a target variable $T$ consisting of a separate elementary event for each joint event $S_{1,2}$. There are several important points to note about the decomposition of TBC. Firstly, due to the symmetry in the probability distribution, each realisation will have the same pointwise decomposition. Secondly, due to the construction of the target, there is an isomorphism (Again, isomorphism should be taken to mean isomorphic probability spaces, e.g., [37], p. 27 or [38], p. 4) between $P(T)$ and $P\left(S_{1}, S_{2}\right)$, and hence the pointwise ambiguity provided by any (individual or joint) predictor event is 0 bit (since given $t$, one knows $s_{1}$ and $s_{2}$ ). Thirdly, the individual predictor events $s_{1}$ and $s_{2}$ each exclude $1 / 2$ of the total probability mass in $P(T)$ and so each provide 1 bit of pointwise specificity; thus, by (23), there is 1 bit of redundant pointwise specificity in each realisation. Fourthly, the joint predictor event $s_{1,2}$ excludes $3 / 4$ of the total probability mass, providing 2 bit of pointwise specificity; hence, by (18), each joint realisation provides 1 bit of pointwise complementary specificity in addition to the 1 bit of redundant pointwise specificity. Finally, putting this together via (22), TBC consists of 1 bit of average redundant information and 1 bit of average complementary information.

Although "surprising" ([5], p. 268), according to the operational interpretation adopted in Section 3.3, two independently distributed predictor variables can share redundant information. That is, since the exclusions induced by $s_{1}$ and $s_{2}$ are the same with respect to the two-event partition $\mathcal{T}^{t}$, the information associated with these exclusions is regarded as being the same. Indeed, this probability distribution highlights the significance of specific reference to the two-event partition in Section 3.3 and Axiom 4. (This can be seen in the probability mass diagram in Figure 7, where the events $S_{1}=0$ and $S_{2}=0$ exclude different elementary target events within the complementary event $0^{c}$ and yet are considered to be the same exclusion with respect to the two-event partition $\mathcal{T}^{0}$.) That these exclusions should be regarded as being the same is discussed further in Appendix A. Now however, there is a need to discuss TBC in terms of Theorem 5 (Target Chain Rule).


Figure 7. Example Tbc. (Top) the probability mass diagrams for the realisation ( $S_{1}=0, S_{2}=0, T=00$ );
(Middle) for each realisation, the PPID using specificity and ambiguity is evaluated (see Figure 4);
(Bottom) the decomposition of XOR yields the same result as $I_{\text {min }}$.

TBC was first considered as a "mechanism" ([6], p.3) where "the wires don't even touch" ([12], p. 167), which merely copies or concatenates $S_{1}$ and $S_{2}$ into a composite target variable $T_{1,2}=\left(T_{1}, T_{2}\right)$ where $T_{1}=S_{1}$ and $T_{2}=S_{2}$. However, using causal mechanisms as a guiding intuition is dubious since different mechanisms can yield isomorphic probability distributions ([44], and references therein). In particular, consider two mechanisms which generate the composite target variables $T_{1,3}=\left(T_{1}, T_{3}\right)$ and $T_{2,3}=\left(T_{2}, T_{3}\right)$ where $T_{3}=S_{1}$ XOR $S_{2}$. As can be seen in Figure 7, both of these mechanisms generate the same (isomorphic) probability distribution $P\left(S_{1}, S_{2}, T\right)$ as the mechanism generating $T_{1,2}$. If an information decomposition is to depend only on the probability distribution $P\left(S_{1}, S_{2}, T\right)$, and no other semantic details such as labelling, then all three mechanisms must yield the same information decomposition-this is not clear from the mechanistic intuition.

Although the decomposition of the various composite target variables must be the same, there is no requirement that the three systems must yield the same decomposition when analysed in terms of the individual components of the composite target variables. Nonetheless, there ought to be a consistency between the decomposition of the composite target variables and the decomposition of the component target variables-i.e., there should be a target chain rule. As shown in Theorem 5, the measures $r_{\text {min }}$ and $R_{\min }$ satisfy the target chain rule, whereas $I_{\min }, \widetilde{U} I, I_{\text {red }}$ and $\mathcal{S}_{\mathrm{VK}}$ do not $[5,7]$. Failing to satisfy the target chain rule can lead to inconsistencies between the composite and component decompositions, depending on the order in which one considers decomposing the information (this is discussed further in Appendix A.3). In particular, Table 2 shows how $\widetilde{U I}, I_{\text {red }}$ and $\mathcal{S}_{\mathrm{VK}}$ all provide the same inconsistent decomposition for TBC when considered in terms of the composite target variable $T_{1,3}$. In contrast, $R_{\min }$ produces a consistent decomposition of $T_{1,3}$. Finally, based on the above isomorphism, consider the following (the proof is deferred to Appendix B.3).

Theorem 6. The target chain rule, identity property and local positivity, cannot be simultaneously satisfied.

Table 2. Shows the decomposition of the quantities in the first row induced by the measures in the first column. For consistency, the decomposition of $I\left(S_{1,2} ; T_{1,3}\right)$ should equal both the sum of the decomposition of $I\left(S_{1,2} ; T_{1}\right)$ and $I\left(S_{1,2} ; T_{3} \mid T_{1}\right)$, and the sum of the decomposition of $I\left(S_{1,2} ; T_{3}\right)$ and $I\left(S_{1,2} ; T_{1} \mid 3\right)$. Note that the decomposition induced by $\widetilde{U I}, I_{\text {red }}$ and $\mathcal{S}_{\mathrm{VK}}$ are not consistent. In contrast, $R_{\text {min }}$ is consistent due to Theorem 5 .

|  | $I\left(S_{1,2} ; T_{1,3}\right)$ | $I\left(S_{1,2} ; T_{1}\right)$ | $I\left(S_{1,2} ; T_{3} \mid T_{1}\right)$ | $I\left(S_{1,2} ; T_{3}\right)$ | $I\left(S_{1,2} ; T_{1} \mid T_{3}\right)$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $\begin{gathered} \widetilde{U I I}, I_{\mathrm{red}}, \\ \mathcal{S}_{\mathrm{VK}} \end{gathered}$ | $U\left(S_{1} \backslash S_{2} \rightarrow T_{1,3}\right)=1$ $U\left(S_{2} \backslash S_{1} \rightarrow T_{1,3}\right)=1$ | $U\left(S_{1} \backslash S_{2} \rightarrow T_{1}\right)=1$ | $U\left(S_{2} \backslash S_{1} \rightarrow T_{3} \mid T_{1}\right)=1$ | $C\left(S_{1}, S_{2} \rightarrow T_{3}\right)=1$ | $R\left(S_{1}, S_{2} \rightarrow T_{1} \mid T_{3}\right)=1$ |
| $R_{\text {min }}$ | $R\left(S_{1}, S_{2} \rightarrow T_{1,3}\right)=1$ $C\left(S_{1}, S_{2} \rightarrow T_{1,3}\right)=1$ | $\begin{aligned} & U\left(S_{2} \backslash S_{1} \rightarrow T_{1}\right)=-1 \\ & R\left(S_{1}, S_{2} \rightarrow T_{1}\right)=1 \\ & C\left(S_{1}, S_{2} \rightarrow T_{1}\right)=1 \end{aligned}$ | $U\left(S_{2} \backslash S_{1} \rightarrow T_{3} \mid T_{1}\right)=1$ | $C\left(S_{1}, S_{2} \rightarrow T_{3}\right)=1$ | $R\left(S_{1}, S_{2} \rightarrow T_{1} \mid T_{3}\right)=1$ |

### 5.6. Summary of Key Properties

The following are the key properties of the PPID using the specificity and ambiguity. Property 1 follows directly from the Definitions 1 and 2. Property 2 follows from Theorems 3 and 4. Property 3 follows from the probability distribution TBC in Section 5.5. Property 4 was discussed in Section 4.2. Property 5 is proved in Theorem 5.

Property 1. When considering the redundancy between the source events $\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k}$, at least one source event $a_{i}$ will provide zero unique specificity, and at least one source event $a_{j}$ will provide zero unique ambiguity. The events $\boldsymbol{a}_{i}$ and $\boldsymbol{a}_{j}$ are not necessarily the same source event.

Property 2. The atoms of partial specificity and partial ambiguity satisfy local positivity, $\pi^{ \pm} \geq 0$. However, upon recombination and averaging, the atoms of partial information do not satisfy local positivity, $\Pi \geq 0$.

Property 3. The decomposition does not satisfy the identity property.
Property 4. The decomposition does not satisfy the target monotonicity property.
Property 5. The decomposition satisfies the target chain rule.

## 6. Conclusions

The partial information decomposition of Williams and Beer [12] provided an intriguing framework for the decomposition of multivariate information. However, it was not long before "serious flaws" ([11], p. 2163) were identified. Firstly, the measure of redundant information $I_{\text {min }}$ failed to distinguish between whether predictor variables provide the same information or merely the same amount of information. Secondly, $I_{\min }$ fails to satisfy the target chain rule, despite this addativity being one of the defining characteristics of information. Notwithstanding the problems, the axiomatic derivation of the redundancy lattice was too elegant to be abandoned and hence several alternate measures were proposed, i.e., $I_{\text {red }}, \widetilde{U I}$ and $\mathcal{S}_{\mathrm{VK}}[6,11,12]$. Nevertheless, as these measures all satisfy the identity property, they cannot produce a non-negative decomposition for an arbitrary number of variables [13]. Furthermore, none of these measures satisfy the target chain rule meaning they produce inconsistent decompositions for multiple target variables. Finally, in spite of satisfying the identity property (which many consider to be desirable), these measures still fail to identify when variables provide the same information, as exemplified by the pointwise unique problem presented in Section 2.

This paper took the axiomatic derivation of the redundancy lattice from PID and applied it to the unsigned entropic components of the pointwise mutual information. This yielded two separate redundancy lattices-the specificity and the ambiguity lattices. Then based upon an operational interpretation of redundancy, measures of pointwise redundant specificity $r_{\text {min }}^{+}$and pointwise redundant
ambiguity $r_{\text {min }}^{-}$were defined. Together with specificity and ambiguity lattices, these measures were used to decompose multivariate information for an arbitrary number of variables. Crucially, upon recombination, the measure $r_{\text {min }}$ satisfies the target chain rule. Furthermore, when applied to PWUNQ, these measures do not result in the pointwise unique problem. In our opinion, this demonstrates that the decomposition is indeed correctly identifying redundant information. However, others will likely disagree with this point given that the measure of redundancy does not satisfy the identity property. According to the identity property, independent variables can never provide the same information. In contrast, according to the operational interpretation adopted in this paper, independent variables can provide the same information if they happen to provide the same exclusions with respect to the two-event target distribution. In any case, the proof of Theorem 6 and the subsequent discussion in Appendix B.3, highlights the difficulties that the identity property introduces when considering the information provided about events in separate target variables. (See further discussion in Appendix A.3).

Our future work with this decomposition will be both theoretical and empirical. Regarding future theoretical work, given that the aim of information decomposition is to derive measures pertaining to sets of random variables, it would be worthwhile to derive the information decomposition from first principles in terms of measure theory. Indeed, such an approach would surely eliminate the semantic arguments (about what it means for information to unique, redundant or complementary), which currently plague the problem domain. Furthermore, this would certainly be a worthwhile exercise before attempting to generalise the information decomposition to continuous random variables. Regarding future empirical work, there are many rich data sets which could be decomposed using this decomposition including financial time-series and neural recordings, e.g., [28,33,34].

Acknowledgments: Joseph T. Lizier was supported through the Australian Research Council DECRA grant DE160100630. We thank Mikhail Prokopenko, Richard Spinney, Michael Wibral, Nathan Harding, Robin Ince, Nils Bertschinger, and Nihat Ay for helpful discussions relating to this manuscript. We also thank the anonymous reviewers for their particularly detailed and helpful feedback.
Author Contributions: C.F. and J.L. conceived the idea; C.F. designed, wrote and analyzed the computational examples; C.F. and J.L. wrote the manuscript.
Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

## Appendix A. Kelly Gambling, Axiom 4, and Tbс

In Section 3.3, it was argued that the information provided by a set of predictor events $s_{1}, \ldots, s_{k}$ about a target event $t$ is the same information if each source event induces the same exclusions with respect to the two-event partition $\mathcal{T}^{t}=\left\{t, t^{c}\right\}$. This was based on the fact that pointwise mutual information does not depend on the apportionment of the exclusions across the set of events which did not occur $t^{c}$. It was argued that since the pointwise mutual information is independent of these differences, the redundant mutual information should also be independent of these differences. This requirement was then integrated into the operational interpretation of Section 3.3 and was later enshrined in the form of Axiom 4. This appendix aims to justify this operational interpretation and argue why redundant information in TBC is not "unreasonably large" ([5], p. 269).

## Appendix A.1. Pointwise Side Information and the Kelly Criterion

Consider a set of horses $\mathcal{T}$ running in a race which can be considered a random variable $T$ with distribution $P(T)$. Say that for each $t \in \mathcal{T}$ a bookmaker offers odds of $o(t)$-for-1, i.e., the bookmaker will pay out $o(t)$ dollars on a $\$ 1$ bet if the horse $t$ wins. Furthermore, say that there is no track take as $\sum_{t \in \mathcal{T}^{1}} / o(t)=1$, and these odds are fair, i.e., $o(t)=1 / p(t)$ for all $t \in \mathcal{T}$ [40]. Let $\boldsymbol{b}(T)$ be the fraction of a gambler's capital bet on each horse $t \in \mathcal{T}$ and assume that the gambler stakes all of their capital on the race, i.e., $\sum_{t \in \mathcal{T}} b(t)=1$.

Now consider an i.i.d. series of these races $T_{1}, T_{2}, \ldots$ such that $P\left(T_{k}\right)=P(T)$ for all $k \in \mathbb{N}$ and let $t_{k} \in \mathcal{T}$ represent the winner of the $k$-th race. Say that the bookmaker offers the same odds on each race and the gambler bets their entire capital on each race. The gambler's capital after $m$ races $D_{m}$ is a random variable which depends on two factors per race: the amount the gambler staked on each race winner $t_{k}$, and the odds offered on each winner $t_{k}$. That is,

$$
\begin{equation*}
D_{m}=\prod_{k=1}^{m} b\left(t_{k}\right) o\left(t_{k}\right), \tag{A1}
\end{equation*}
$$

where monetary units $\$$ have been chosen such that $D_{0}=\$ 1$. The gambler's wealth grows (or shrinks) exponentially, i.e.,

$$
\begin{equation*}
D_{m}=2^{m W(\boldsymbol{b}, T)} \tag{A2}
\end{equation*}
$$

where

$$
\begin{equation*}
W(\boldsymbol{b}, T)=\frac{1}{m} \log D_{m}=\frac{1}{m} \sum_{k=1}^{m} \log b\left(t_{k}\right) o\left(t_{k}\right)=\mathrm{E}\left[\log b\left(t_{k}\right) o\left(t_{k}\right)\right] \tag{A3}
\end{equation*}
$$

is the doubling rate of the gambler's wealth using a betting strategy $\boldsymbol{b}(T)$. Here, the last equality is by the weak law of large numbers for large $m$.

Any reasonable gambler would aim to use an optimal strategy $\boldsymbol{b}^{*}(T)$ which maximises the doubling rate $W(\boldsymbol{b}, T)$. Kelly $[40,43]$ proved that the optimal doubling rate is given by

$$
\begin{equation*}
W^{*}(T)=\max _{\boldsymbol{b}} W(\boldsymbol{b}, T)=\mathrm{E}\left[\log b^{*}\left(t_{k}\right) o\left(t_{k}\right)\right] \tag{A4}
\end{equation*}
$$

and is achieved by using the proportional gambling scheme $\boldsymbol{b}^{*}(T)=P(T)$. When the race $T_{k}$ occurs and the horse $t_{k}$ wins, the gambler will receive a payout of $b^{*}\left(t^{k}\right) o\left(t^{k}\right)=\$ 1$, i.e., the gambler receives their stake back regardless of the outcome. In the face of fair odds, the proportional Kelly betting scheme is the optimal strategy-non-terminating repeated betting with any other strategy will result in losses.

Now consider a gambler with access to a private wire $S$ which provides (potentially useful) side information about the upcoming race. Say that these messages are selected from the set $\mathcal{S}$, and that the gambler receives the message $s_{k}$ before the race $T_{k}$. Kelly $[40,43]$ showed that the optimal doubling rate in the presence of this side information is given by

$$
\begin{equation*}
W^{*}(T \mid S)=\max _{b} W(\boldsymbol{b}, T \mid S)=\mathrm{E}\left[\log b^{*}\left(t_{k} \mid s_{k}\right) o\left(t_{k}\right)\right] \tag{A5}
\end{equation*}
$$

and is achieved by using the conditional proportional gambling scheme $\boldsymbol{b}^{*}\left(T \mid s_{k}\right)=P\left(T \mid s_{k}\right)$. Both the proportional gambling scheme $b^{*}(T)$ and the conditional proportional gambling scheme $b^{*}(T \mid S)$ are based upon the Kelly criterion whereby bets are apportioned according to the best estimation of the outcome available. The financial value of the private wire to a gambler can be ascertained by comparing their doubling rate of the gambler with access to the side wire to that of a gambler with no side information, i.e.,

$$
\begin{align*}
\Delta W=W^{*}(T \mid S)-W^{*}(T) & =\mathrm{E}\left[\log b^{*}\left(t_{k} \mid s_{k}\right) o\left(t_{k}\right)\right]-\mathrm{E}\left[\log b^{*}\left(t_{k}\right) o\left(t_{k}\right)\right] \\
& =\mathrm{E}\left[i\left(s_{k} ; t_{k}\right)\right]=I(S ; T) \tag{A6}
\end{align*}
$$

This important result due to Kelly [40] equates the increase in the doubling rate $\Delta W$ due to the presence of side information, with the mutual information between the private wire $S$ and the horse race $T$. If on average, the gambler receives 1 bit of information from their private wire, then on average the gambler can expect to double their money per race. Furthermore, as one would expect, independent side information does not increase the doubling rate.

With no side information, the Kelly gambler always received their original stake back from the bookmaker. However, this is not true for the Kelly gambler with side information. Although their doubling rate is greater than or equal to that of the gambler with no side information, this is only true on average. Before the race $T_{k}$, the gambler receives the private wire message $s_{k}$ and then, the horse $t_{k}$ wins the race. From (A6), one can see that the return $\Delta w_{k}$ for the $k$-th race is given by the pointwise mutual information,

$$
\begin{equation*}
\Delta w=i\left(s_{k} ; t_{k}\right) \tag{A7}
\end{equation*}
$$

Hence, just like the pointwise mutual information, the per race return can be positive or negative: if it is positive, the gambler will make a profit; if it is negative, the gambler will sustain a loss. Despite the potential for pointwise loses, the average return (i.e., the doubling rate) is, just like the average mutual information, non-negative-and indeed, is optimal. Furthermore, while a Kelly gambler with side information can lose money on any single race, they can never actually go bust. The Kelly gambler with side information $s$ still hedges their risk by placing bets on all horses with a non-zero probability of winning according to their side information, i.e., according to $P\left(T \mid s_{k}\right)$. The only reason they would fail to place a bet on a horse is if their side information completely precludes any possibility of that horse winning. That is, a Kelly gambler with side information will never fall foul of gambler's ruin.

## Appendix A.2. Justification of Axiom 4 and Redundant Information in TBC

Consider TBC semantically described in terms of a horse race. That is, consider a four horse race $T$ where each horse has an equiprobable chance of winning, and consider the binary variables $T_{1}, T_{2}$, and $T_{3}$ which represent the following, respectively: the colour of the horse, black 0 or white 1 ; the sex of the jockey, female 0 or male 1 ; and the colour of the jockey's jersey, red 0 or green 1 . Say that the four horses have the following attributes:

Horse 0 is a black horse $T_{1}=0$, ridden by a female jockey $T_{2}=0$, who is wearing a red jersey $T_{3}=0$.
Horse 1 is a black horse $T_{1}=0$, ridden by a male jockey $T_{2}=1$, who is wearing a green jersey $T_{3}=1$.
Horse 2 is a white horse $T_{1}=1$, ridden by a female jockey $T_{2}=0$, who is wearing a green jersey $T_{3}=1$. Horse 3 is a white horse $T_{1}=1$, ridden by a male jockey $T_{2}=1$, who is wearing a red jersey $T_{3}=0$.

There are two important points to note. Firstly, the horses in the race $T$ could also be uniquely described in terms of the composite binary variables $T_{1,2}, T_{1,3}$ or $T_{2,3}$. Secondly, if one knows $T_{1}$ and $T_{2}$ then one knows $T_{3}$ (which can be represented by the relationship $T_{3}=T_{1}$ XOR $T_{2}$ ). Finally, consider private wires $S_{1}$ and $S_{2}$ which independently provide the colour of the horse and the colour of the jockey's jersey (respectively) before the upcoming race, i.e., $S_{1}=T_{1}$ and $S_{2}=T_{2}$.

Now say a bookmaker offers fair odds of 4 -for- 1 on each horse in the race $T$. Consider two gamblers who each have access to one of $S_{1}$ and $S_{2}$. Before each race, the two gamblers receive their respective private wire messages and place their bets according to the Kelly strategy. This means that each gambler lays half of their, say $\$ 1$, stake on each of their two respective non-excluded horses: unknowingly, both of the gamblers have placed a bet on the soon-to-be race winner, and each gambler has placed a distinct bet on one of the two soon-to-be losers. The only horse neither has bet upon is also a soon-to-be loser. (See [5] for a related description of TBC in term of the game-theoretic notions of shared and common knowledge). After the race, the bookmaker pays out $\$ 2$ dollars to each gambler: both have doubled their money. This happens because both of the gamblers had one bit of 1 bit of information about the race, i.e., pointwise mutual information. In particular, both gamblers improved their probability of predicting the eventual race winner. It did not matter, in any way, that the gamblers had each laid distinct bets on one of the three eventual race losers. The fact that they laid different bets on the horses which did not win, made no difference to their winnings. The apportionment of the exclusions across the set of events which did not occur, makes no difference to the pointwise mutual information. With respect to what occurred (i.e., with respect to which horse won), the fact the that they excluded different losers is only semantic. When it came to predicting the would-be-winner,
both gamblers had the same predictive power; they both had the same freedom of choice with regards to selecting what would turn out to be the eventual race winner-they had the same information. It is for this reason that this information should be regarded as redundant information, regardless of the independence of the information sources. Hence, the introduction of both the operational interpretation of redundancy in Section 3.3 and Axiom 4 in Section 4.2.

Now consider a third gambler who has access to both private wires $S_{1}$ and $S_{2}$, i.e., $S_{1,2}$. Before the race, this gambler receives both private wire messages which, in total, precludes three of the horses from winning. This gambler then places the entirety of their $\$ 1$ stake on the remaining horse which is sure to win. After the race, the bookmaker pays out $\$ 4$ : this gambler has quadrupled their money as they had 2 bit of pointwise mutual information about the race. Having both private wire messages simultaneously gave this gambler a 1 bit informational edge over the two gamblers with access to a single side wire. While each of the singleton gamblers had 1 bit of independent information, the only way one could profit from the independence of this information is by having both pieces of information simultaneously-this makes this 1 bit of information complementary. Although this may seem "palpably strange" ([12], p. 167), it is not so strange when from the following perspective: the only way to exploit two pieces of independent information is by having both pieces together simultaneously.

## Appendix A.3. Accumulator Betting and the Target Chain Rule

Say that in addition to the 4 -for- 1 odds offered on the race $T$, the bookmaker also offers fair odds of 2-for- 1 on each of the binary variables $T_{1}, T_{2}$ and $T_{3}$. Now, in addition to being able to directly gamble on the race $T$, one could indirectly gamble on $T$ by placing a so-called accumulator bet on any pair of $T_{1}, T_{2}$ and $T_{3}$. An accumulator is a series of chained bets whereby any return from one bet is automatically staked on the next bet; if any bet in the chain is lost then the entire chain is lost. For example, a gambler could place 4 -for- 1 bet on horse 0 by placing the following accumulator bet: a 2 -for 1 bet on a black horse winning that chains into a 2 -for- 1 bet on the winning jockey being female (or equivalently, vice versa). In effect, these accumulators enable a gambler to bet on $T$ by instead placing a chained bet on the independent component variables within the (equivalent) joint variables $T_{1,2}, T_{1,3}$ and $T_{2,3}$. Now consider again the three gamblers from the prior section, i.e., the two gamblers who each have a private wire $S_{1}$ and $S_{2}$, and the third gamble who has access to $S_{1,2}$. Say that they must each place a, say $\$ 1$, accumulator bet on $T_{1,3}$-what should each gambler do according to the Kelly criterion?

For the sake of clarity, consider only the realisation where the horse $T=0$ subsequently wins (due to the symmetry, the analysis is equivalent for all realisations). First consider the accumulator whereby the gamblers first bet on the colour of the winning horse $T_{1}$, which chains into a bet on the colour of the winning jockey's jersey $T_{3}$. Suppose that the private wire $S_{1}$ communicates that the winning horse will be black, while the private wire $S_{2}$ communicates that the winning horse will be ridden by a female jockey, i.e., $S_{1}=0$ and $S_{2}=0$. Following to the Kelly strategy, the gambler with access to $S_{1}=0$ takes out two $\$ 0.5$ accumulator bets. Both of these accumulators feature the same initial bet on the winning horse being black since $T_{1}=S_{1}=0$. Hence both bets return $\$ 1$ each which become the stake on the next bet in each accumulator. This gambler knows nothing about the colour of the jockey's jersey $T_{3}$. As such, one accumulator chains into a bet on the winning jersey being red $T_{3}=0$, while the other chains into a bet on it being green $T_{3}=1$. When the horse $T=0$ wins, the stake bet on the green jersey is lost while bet on red jersey pays out $\$ 2$. This gambler had 1 bit of side information and so doubled their money. Now consider the gambler with private wire $S_{2}$, who knows nothing about $T_{1}$ or $T_{3}$ individually. Nonetheless, this gambler knows that the winner must be a female jockey $T_{2}=0$. As such, this gambler knows that if a black horse $T_{1}=0$ wins then its jockey must be wearing a red jersey $T_{3}=0$, or if a white horse $T_{1}=0$ wins then its jockey must be wearing a green jersey $T_{3}=1$ (since $T_{3}=T_{1}$ XOR $T_{2}$ ). Thus this gambler can also utilise the Kelly strategy to place the following two $\$ 0.5$ accumulator bets: the first accumulator bets on the winning horse being black $T_{1}=0$ and then chains into a bet on the winner's jersey being red $T_{3}=0$, while the second accumulator bets on the winning horse being white $T_{1}=1$ and then chains into a bet on the winner's jersey being green
$T_{3}=1$. When the horse $T=0$ wins, the first accumulator pays out $\$ 2$, while the second accumulator is be lost. Hence, this gambler also doubles their money and so also had 1 bit of side information. Finally, consider the gambler with access to both private wires $S_{1,3}$, who can place an accumulator on the black horse $T_{1}=0$ winning chaining into a bet on the winning jockey wearing red $T_{3}=0$. This gambler can quadruple their stake, and so must possess 2 bit of side information.

Each of the three gamblers have the same final return regardless of whether the gamblers are betting on the variable $T$, or placing accumulator bets on the variables $T_{1,2}, T_{1,3}$ or $T_{2,3}$. However, the paths to the final result differs between the gamblers, reflecting the difference between the information the each gambler had about the sub-variables $T_{1}, T_{2}$ or $T_{3}$. Given the result of Kelly [40], the proposed information decomposition should reflect these differences, but yet still arrive at the same result-in other words, the information decomposition should satisfy a target chain rule. This is clear if the Kelly interpretation of information is to remain as a "duality" ([43], p. 159) in information theory.

## Appendix B. Supporting Proofs and Further Details

This appendix contains many of the important theorems and proofs relating to PPID using specificity and ambiguity.

## Appendix B.1. Deriving the Specificity and Ambiguity Lattices from Axioms 1-4

The following section is based directly on the original work of Williams and Beer [1,2]. The difference is that we now consider sources events $\boldsymbol{a}_{i}$ rather than sources $\boldsymbol{A}_{i}$.

Proposition A1. Both $i_{\cap}^{+}$and $i_{\cap}^{-}$are non-negative.
Proof. Since $\varnothing \subseteq \boldsymbol{a}_{i}$ for any $\boldsymbol{a}_{i}$, Axioms 2 and 3 imply

$$
\begin{align*}
& i_{\cap}^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t\right) \geq i_{\cap}^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k}, \varnothing \rightarrow t\right)=i_{\cap}^{+}(\varnothing \rightarrow t)=h(\varnothing)=0  \tag{A8}\\
& i_{\cap}^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t\right) \geq i_{\cap}^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k}, \varnothing \rightarrow t\right)=i_{\cap}^{-}(\varnothing \rightarrow t)=h(\varnothing \mid t)=0 \tag{A9}
\end{align*}
$$

Hence, both $i_{\cap}^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t\right)$ and $i_{\cap}^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t\right)$ are non-negative.
Proposition A2. Both $i_{\cap}^{+}$and $i_{\cap}^{-}$are bounded from above by the specificity and the ambiguity from any single source event, respectively.

Proof. For any single source $\boldsymbol{a}_{i}$, Axioms 2 and 3 yield

$$
\begin{align*}
& h\left(a_{i}\right)=i_{\cap}^{+}\left(\boldsymbol{a}_{i} \rightarrow t\right)=i_{\cap}^{+}\left(\boldsymbol{a}_{i}, \boldsymbol{a}_{i} \rightarrow t\right) \geq i_{\cap}^{+}\left(\boldsymbol{a}_{i}, \ldots \rightarrow t\right),  \tag{A10}\\
& h\left(a_{i} \mid t\right)=i_{\cap}^{-}\left(\boldsymbol{a}_{i} \rightarrow t\right)=i_{\cap}^{-}\left(\boldsymbol{a}_{i}, \boldsymbol{a}_{i} \rightarrow t\right) \geq i_{\cap}^{+}\left(\boldsymbol{a}_{i}, \ldots \rightarrow t\right), \tag{A11}
\end{align*}
$$

as required.
In keeping with Williams and Beer's approach [1,2], consider all of the distinct ways in which a collection of source events $\boldsymbol{a}=\left\{\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k}\right\}$ could contribute redundant information. Thus far we have assumed that the redundancy measure can be applied to any collection of source events, i.e., $\mathscr{P}_{1}(\boldsymbol{a})$ where $\mathscr{P}_{1}$ denotes the power set with the empty set removed. Recall that the sources events are themselves collections of predictor events, i.e., $\mathscr{P}_{1}(\boldsymbol{s})$. That is, we can apply both $i_{\cap}^{+}$and $i_{\cap}^{-}$to elements of $\mathscr{P}_{1}\left(\mathscr{P}_{1}(s)\right)$. However, this can be greatly reduced using Axiom 2 which states that if $\boldsymbol{a}_{i} \subseteq \boldsymbol{a}_{j}$, then

$$
\begin{align*}
i_{\cap}^{+}\left(\boldsymbol{a}_{j}, \boldsymbol{a}_{i}, \ldots \rightarrow t\right) & =i_{\cap}^{+}\left(\boldsymbol{a}_{i}, \ldots \rightarrow t\right)  \tag{A12}\\
i_{\cap}^{-}\left(\boldsymbol{a}_{j}, \boldsymbol{a}_{i}, \ldots \rightarrow t\right) & =i_{\cap}^{-}\left(\boldsymbol{a}_{i}, \ldots \rightarrow t\right) \tag{A13}
\end{align*}
$$

Hence, one need only consider the collection of source events such that no source event is a superset of any other in order,

$$
\begin{equation*}
\mathscr{A}(\boldsymbol{s})=\left\{\alpha \in \mathscr{P}_{1}\left(\mathscr{P}_{1}(\boldsymbol{s})\right) \mid \forall \boldsymbol{a}_{i}, \boldsymbol{a}_{j} \in \alpha, \boldsymbol{a}_{i} \not \subset \boldsymbol{a}_{j}\right\} . \tag{A14}
\end{equation*}
$$

This collection $\mathscr{A}(s)$ captures all the distinct ways in the source events could provide redundant information.

As per Williams and Beer's PID, this set of source events $\mathscr{A}(s)$ is structured. Consider two sets of source events $\alpha, \beta \in \mathscr{A}(s)$. If for every source event $\boldsymbol{b} \in \beta$ there exists a source event $\boldsymbol{a} \in \alpha$ such that $\boldsymbol{a} \subseteq \boldsymbol{b}$, then all of the redundant specificity and ambiguity shared by $\boldsymbol{b} \in \beta$ must include any redundant specificity and ambiguity shared by $a \in \alpha$. Hence, a partial order $\preceq$ can be defined over the elements of the domain $\mathscr{A}(\boldsymbol{s})$ such that any collection of predictors event coalitions precedes another if and only if the latter provides any information the former provides,

$$
\begin{equation*}
\forall \alpha, \beta \in \mathscr{A}(\boldsymbol{s}),(\alpha \preceq \beta \Longleftrightarrow \forall \boldsymbol{b} \in \beta, \exists \boldsymbol{a} \in \alpha \mid \boldsymbol{a} \subseteq \boldsymbol{b}) . \tag{A15}
\end{equation*}
$$

Applying this partial ordering to the elements of the domain $\mathscr{A}(s)$ produces a lattice which has the same structure as the redundancy lattice from PID, i.e., the structure of the sources events here is the same as the structure of the sources in PID. (Figure 3 depicts this structure for the case of 2 and 3 predictor variables.) Applying $i_{\cap}^{+}$to these sources events yields a specificity lattice while applying $i_{\cap}^{-}$yields an ambiguity lattice.

Similar to $I_{\cap}$ in PID, the redundancy measures $i_{\cap}^{+}$or $i_{\cap}^{-}$can be thought of as a cumulative information functions which integrate the specificity or ambiguity uniquely contributed by each node as one moves up each lattice. In order in evaluate the unique contribution of specificity and ambiguity from each node in the lattice, consider the Möbius inverse $[45,46]$ of $i_{\cap}^{+}$and $i_{\cap}^{-}$. That is, the specificity and ambiguity of a node $\alpha$ is given by

$$
\begin{equation*}
i_{\cap}^{ \pm}(\alpha \rightarrow t)=\sum_{\beta \preceq \alpha} i_{\cap}^{ \pm}(\beta \rightarrow t) \quad \forall \alpha, \beta \in \mathscr{A}(s) . \tag{A16}
\end{equation*}
$$

Thus the unique contributions of partial specificity $i_{\partial}^{+}$and partial ambiguity $i_{\partial}^{-}$from each node can be calculated recursively from the bottom-up, i.e.,

$$
\begin{equation*}
i_{\partial}^{ \pm}(\alpha \rightarrow t)=i_{\cap}^{ \pm}(\alpha \rightarrow t)-\sum_{\beta \prec \alpha} i_{\partial}^{ \pm}(\beta \rightarrow t) \tag{A17}
\end{equation*}
$$

Theorem A1. Based on the principle of inclusion-exclusion, we have the following closed-from expression for the partial specificity and partial ambiguity,

$$
\begin{equation*}
i_{\partial}^{ \pm}(\alpha \rightarrow t)=i_{\cap}^{ \pm}(\alpha \rightarrow t)-\sum_{\varnothing \neq \gamma \subseteq \alpha^{-}}(-1)^{|\gamma|-1} i_{\cap}^{ \pm}(\bigwedge \gamma \rightarrow t) \tag{A18}
\end{equation*}
$$

Proof. For $\mathscr{B} \subseteq \mathscr{A}(s)$, define the sub-addative function $f^{ \pm}(\mathscr{B})=\sum_{\beta \in \mathscr{B}}=i^{ \pm}(\beta \rightarrow t)$. From (A16), we get that $i_{\cap}^{ \pm}(\alpha \rightarrow t)=f^{ \pm}(\downarrow \alpha)$ and

$$
\begin{equation*}
i_{\partial}^{ \pm}(\alpha \rightarrow t)=f^{ \pm}(\downarrow \alpha)-f^{ \pm}(\downarrow \alpha)=f^{ \pm}(\downarrow \alpha)-f^{ \pm}\left(\bigcup_{\beta \in \alpha^{-}} \downarrow \beta\right) \tag{A19}
\end{equation*}
$$

By the principle of inclusion-exclusion (e.g., see [46], p. 195) we get that

$$
\begin{equation*}
i_{\partial}^{ \pm}(\alpha \rightarrow t)=f^{ \pm}(\downarrow \alpha)-\sum_{\varnothing \neq \gamma \subseteq \alpha^{-}}(-1)^{|\gamma|-1} f^{ \pm}\left(\bigcap_{\beta \in \gamma} \beta\right) \tag{A20}
\end{equation*}
$$

For any lattice $L$ and $A \subseteq L$, we have that $\cap_{a \in A} \downarrow a=\downarrow(\wedge A)$ (see [47], p. 57) thus

$$
\begin{align*}
i_{\partial}^{ \pm}(\alpha \rightarrow t) & =f^{ \pm}(\downarrow \alpha)-\sum_{\varnothing \neq \gamma \subseteq \alpha^{-}}(-1)^{|\gamma|-1} f^{ \pm}(\bigwedge \gamma) \\
& =f^{ \pm}(\downarrow \alpha)-\sum_{\varnothing \neq \gamma \subseteq \alpha^{-}}(-1)^{|\gamma|-1} i^{ \pm}(\bigwedge \gamma \rightarrow t) \tag{A21}
\end{align*}
$$

as required.
Similarly to PID, the specificity and ambiguity lattices provide a structure for information decomposition-unique evaluation requires a separate definition of redundancy. However, unlike PID (or even PPID), this evaluation requires both a definition of pointwise redundant specificity and pointwise redundant ambiguity.

## Appendix B.2. Redundancy Measures on the Lattices

In Section 4.2, Definitions 1 and 2 provided the require measures. This section will prove some of the key properties of these measures when they are applies to the lattices derived in the previous section. The correspondence with the approach taken by Williams and Beer [1,2] continues in this section. However, sources events $\boldsymbol{a}_{i}$ are used in place of sources $A_{i}$ and the measures $r_{\text {min }}^{ \pm}$are used in place of $I_{\min }$. Note that the basic concepts from lattice theory and the notion used here are the same as found in ([1], Appendix B).

Theorem 1. The definitions of $r_{\text {min }}^{+}$and $r_{\min }^{-}$satisfy Axioms 1-4.
Proof. Axioms 1, 3 and 4 follow trivially from the basic properties of the minimum. The main statement of Axiom 2 also immediately follows from the properties of the minimum; however, there is a need to verify the equality condition. As such, consider $a_{k}$ such that $a_{k} \supseteq \boldsymbol{a}_{i}$ for some $\boldsymbol{a}_{i} \in$ $\left\{\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k-1}\right\}$. From Postulate 4, we have that $h\left(\boldsymbol{a}_{k}\right) \geq h\left(\boldsymbol{a}_{i}\right)$ and hence that $\min _{\boldsymbol{a}_{j} \in\left\{\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k}\right\}} h\left(\boldsymbol{a}_{j}\right)=$ $\min _{\boldsymbol{a}_{j} \in\left\{\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k-1}\right\}} h\left(\boldsymbol{a}_{j}\right)$, as required for $r_{\min }^{+}$. Mutatis mutandis, similar follows for $r_{\min }^{-}$.

Theorem 2. The redundancy measures $r_{\min }^{+}$and $r_{\min }^{-}$increase monotonically on the $\langle\mathscr{A}(\boldsymbol{s}), \preceq\rangle$.
The proof of this theorem will require the following lemma.
Lemma A1. The specificity and ambiguity $i^{ \pm}(\boldsymbol{a} \rightarrow t)$ are increasing functions on the lattice $\left\langle\mathscr{P}_{1}(\boldsymbol{s}), \subseteq\right\rangle$
Proof. Follows trivially from Postulate 4.
Proof of Theorem 2. Assume there exists $\alpha, \beta \in \mathscr{A}(s)$ such that $\alpha \prec \beta$ and $r_{\text {min }}^{ \pm}(\beta \rightarrow t)<r_{\text {min }}^{ \pm}(\alpha \rightarrow t)$. By definition, i.e., (23) and (24), there exists $\boldsymbol{b} \in \beta$ such that $i^{ \pm}(\boldsymbol{b} \rightarrow t)<i^{ \pm}(\boldsymbol{a} \rightarrow t)$ for all $\boldsymbol{a} \in \alpha$. Hence, by Lemma A1, there does not exist $\boldsymbol{a} \in \alpha$ such that $\boldsymbol{a} \subseteq \boldsymbol{b}$. However, by assumption $\alpha \prec \beta$ and hence there exists $\boldsymbol{a} \in \alpha$ such that $\boldsymbol{a} \subseteq \boldsymbol{b}$, which is a contradiction.

Theorem A2. When using $r_{\text {min }}^{ \pm}$in place of the general redundancy measures $i_{\cap}^{ \pm}$, we have the following closed-from expression for the partial specificity $\pi^{+}$and partial ambiguity $\pi^{-}$,

$$
\begin{equation*}
\pi^{ \pm}(\alpha \rightarrow t)=r_{\min }^{ \pm}(\alpha \rightarrow t)-\max _{\beta \in \alpha^{-}} \min _{b \in \beta} i^{ \pm}(\boldsymbol{b} \rightarrow t) \tag{A22}
\end{equation*}
$$

Proof. Let $i_{\cap}^{+}=r_{\text {min }}^{+}$and $i_{\cap}^{-}=r_{\text {min }}^{-}$in the general closed form expression for $i_{\partial}^{ \pm}$in Theorem A1,

$$
\begin{equation*}
\pi^{ \pm}(\alpha \rightarrow t)=r_{\min }^{ \pm}(\alpha \rightarrow t)-\sum_{\varnothing \neq \gamma \subseteq \alpha^{-}}(-1)^{|\gamma|-1} \min _{b \in \Lambda \gamma} i^{ \pm}(\boldsymbol{b} \rightarrow t) \tag{A23}
\end{equation*}
$$

Since $\alpha \wedge \beta=\underline{\alpha \cup \beta}$ (see [1], Equation (23)), and by Postulate 4, we have that

$$
\begin{equation*}
\pi^{ \pm}(\alpha \rightarrow t)=r_{\min }^{ \pm}(\alpha \rightarrow t)-\sum_{\varnothing \neq \gamma \subseteq \alpha^{-}}(-1)^{|\gamma|-1} \min _{\beta \in \gamma} \min _{\boldsymbol{b} \in \beta} i^{ \pm}(\boldsymbol{b} \rightarrow t) \tag{A24}
\end{equation*}
$$

By the maximum-minimums identity (see [48]), we have that, $\max \alpha^{-}=\sum_{\varnothing \neq \gamma \subseteq \alpha^{-}}(-1)^{|\gamma|-1} \min \gamma$, and hence

$$
\begin{equation*}
\pi^{ \pm}(\alpha \rightarrow t)=r_{\min }^{ \pm}(\alpha \rightarrow t)-\max _{\beta \in \alpha^{-}} \min _{b \in \beta} i^{ \pm}(\alpha \rightarrow t) \tag{A25}
\end{equation*}
$$

as required.
Theorem 3. The atoms of partial specificity $\pi^{+}$and partial ambiguity $\pi^{-}$evaluated using the measures $r_{\text {min }}^{+}$ and $r_{\min }^{-}$on the specificity and ambiguity lattices (respectively), are non-negative.

Proof. It $\alpha=\perp$, the $\pi^{ \pm}(\alpha \rightarrow t)=r_{\text {min }}^{ \pm} \geq 0$ by the non-negativity of entropy. If $\alpha \neq \perp$, assume there exists $\alpha \in \mathscr{A}(s) \backslash\{\perp\}$ such that $\pi^{ \pm}(\alpha \rightarrow t)<0$. By Theorem A2,

$$
\begin{equation*}
\pi^{ \pm}(\alpha \rightarrow t)=\min _{\boldsymbol{a} \in \alpha} i^{ \pm}(\boldsymbol{a} \rightarrow t)-\max _{\beta \in \alpha^{-}} \min _{\boldsymbol{b} \in \beta} i^{ \pm}(\boldsymbol{b} \rightarrow t) \tag{A26}
\end{equation*}
$$

From this it can be seen that there must exist $\beta \in \alpha^{-}$such that for all $\boldsymbol{b} \in \beta$, we have that $i^{ \pm}(\boldsymbol{a} \rightarrow t)<i^{ \pm}(\boldsymbol{b} \rightarrow t)$ for some $\boldsymbol{a} \in \alpha$. By Postulate 4 there does not exist $\boldsymbol{b} \in \beta$ such that $\boldsymbol{b} \subset \boldsymbol{a}$. However, since by definition, $\beta \prec \alpha$ there exists $\boldsymbol{b} \in \beta$ such that $\boldsymbol{b} \subset \boldsymbol{a}$, which is a contradiction.

Theorem 4. The atoms of partial average information $\Pi$ evaluated by recombining and averaging $\pi^{ \pm}$are not non-negative.

Proof. The proof is by the counter-example using RDNERR.

## Appendix B.3. Target Chain Rule

By using the appropriate conditional probabilities in Definitions 1 and 2, one can easily obtain the conditional pointwise redundant specificity,

$$
\begin{equation*}
r_{\min }^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1} \mid t_{2}\right)=\min _{a_{i}} h\left(\boldsymbol{a}_{i} \mid t_{2}\right) \tag{A27}
\end{equation*}
$$

or the conditional pointwise redundant ambiguity,

$$
\begin{equation*}
r_{\min }^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1} \mid t_{2}\right)=\min _{\boldsymbol{a}_{j}} h\left(\boldsymbol{a}_{j} \mid t_{1,2}\right) \tag{A28}
\end{equation*}
$$

As per (21) these could be recombined, e.g., via (21), to obtain the conditional redundant information,

$$
\begin{equation*}
r_{\min }\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1} \mid t_{2}\right)=r_{\min }^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1} \mid t_{2}\right)-r_{\min }^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1} \mid t_{2}\right) \tag{A29}
\end{equation*}
$$

The relationship between the regular forms and the conditional forms of the redundant specificity and redundant ambiguity has some important consequences.

Proposition A3. The conditional pointwise redundant specificity provided by $\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k}$ about $t_{1}$ given $t_{2}$ is equal to pointwise redundant ambiguity provided by $\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k}$ about $t_{2}$ with the conditioned variable,

$$
\begin{equation*}
r_{\text {min }}^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1} \mid t_{2}\right)=r_{\text {min }}^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{2}\right) \tag{A30}
\end{equation*}
$$

Proof. By (24) and (A27).

Proposition A4. The pointwise redundant specificity provided by $\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k}$ is independent of the target event and even the target variable itself,

$$
\begin{equation*}
r_{\text {min }}^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1}\right)=r_{\text {min }}^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{2}\right) \quad \forall t_{1}, t_{2}, T_{1}, T_{2} \tag{A31}
\end{equation*}
$$

Proof. By inspection of (23).
Proposition A5. The conditional pointwise redundant ambiguity provided by $\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k}$ about $t_{1}$ given $t_{2}$ is equal to the pointwise redundant ambiguity provided by $a_{1}, \ldots, a_{k}$ about $t_{1,2}$,

$$
\begin{equation*}
r_{\text {min }}^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1} \mid t_{2}\right)=r_{\text {min }}^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1,2}\right) \tag{A32}
\end{equation*}
$$

Proof. By (24) and (A28).
Note that specificity itself is not a function of the target event or variable. Hence, all of the target dependency is bound up in the ambiguity. Now consider the following.

Theorem 5 (Pointwise Target Chain Rule). Given the joint target realisation $t_{1,2}$, the pointwise redundant information $r_{\text {min }}$ satisfies the following chain rule,

$$
\begin{align*}
r_{\min }\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1,2}\right) & =r_{\min }\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1}\right)+r_{\min }\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{2} \mid t_{1}\right)  \tag{25}\\
& =r_{\min }\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{2}\right)+r_{\min }\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1} \mid t_{2}\right)
\end{align*}
$$

Proof. Starting from $r_{\text {min }}$, by Corollary A4 and Corollary A5 we get that

$$
\begin{align*}
r_{\min }\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1,2}\right) & =r_{\min }^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1,2}\right)-r_{\min }^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1,2}\right)  \tag{A33}\\
& =r_{\min }^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1}\right)-r_{\min }^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{2} \mid t_{1}\right)
\end{align*}
$$

Then, by Corollary A3 we get that

$$
\begin{align*}
r_{\min }\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1,2}\right)= & r_{\min }^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1}\right)-r_{\min }^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1}\right) \\
& \quad+r_{\min }^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1}\right)-r_{\min }^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{2} \mid t_{1}\right), \\
= & r_{\min }^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1}\right)-r_{\min }^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1}\right) \\
& \quad+r_{\min }^{+}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{2} \mid t_{1}\right)-r_{\min }^{-}\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{2} \mid t_{1}\right), \\
= & r_{\min }\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{1}\right)+r_{\min }\left(\boldsymbol{a}_{1}, \ldots, \boldsymbol{a}_{k} \rightarrow t_{2} \mid t_{1}\right), \tag{A34}
\end{align*}
$$

as required for the first equality in (25). Mutatis mutandis, we obtain the second equality in (25).
Theorem 6. The target chain rule, identity property and local positivity, cannot be simultaneously satisfied.
Proof. Consider the probability distribution TBC, and in particular, the isomorphic probability distributions $P\left(T_{1,2}\right)$ and $P\left(T_{1,3}\right)$. By the identity property,

$$
\begin{equation*}
U\left(S_{1} \backslash S_{2} \rightarrow T_{1,2}\right)=1 \mathrm{bit}, \quad U\left(S_{2} \backslash S_{1} \rightarrow T_{1,2}\right)=1 \mathrm{bit} \tag{A35}
\end{equation*}
$$

and hence, $R\left(S_{1}, S_{2} \rightarrow T_{1,2}\right)=0$ bit. On the other hand, by local positivity,

$$
\begin{equation*}
C\left(S_{1}, S_{2} \rightarrow T_{3}\right)=1 \text { bit, } \quad R\left(S_{1}, S_{2} \rightarrow T_{1} \mid T_{3}\right)=1 \mathrm{bit} \tag{A36}
\end{equation*}
$$

Then by the target chain rule,

$$
\begin{equation*}
C\left(S_{1}, S_{2} \rightarrow T_{1,3}\right)=1 \text { bit } \quad R\left(S_{1}, S_{2} \rightarrow T_{1,3}\right)=1 \text { bit } \tag{A37}
\end{equation*}
$$

Finally, since $P\left(T_{1,2}\right)$ is isomorphic to $P\left(T_{1,3}\right)$ we have that, $R\left(S_{1}, S_{2} \rightarrow T_{1,3}\right)=R\left(S_{1}, S_{2} \rightarrow T_{1,2}\right)$, which is a contradiction.

Theorem 6 can be informally generalised as follows: it is not possible to simultaneously satisfy the target chain rule, the identity property, and have only $C\left(S_{1}, S_{2} \rightarrow T\right)=1$ bit in the probability distribution XOR without having negative (average) PI atoms in probability distributions where there is no ambiguity from any source. To see this, again consider decomposing the isomorphic probability distributions $P\left(T_{1,2}\right)$ and $P\left(T_{1,3}\right)$. In line with (A35), decomposing $T_{1,2}$ via the identity property yields $C\left(S_{1}, S_{2} \rightarrow T_{1,2}\right)=0$ bit. On the other hand, decomposing $T_{1,3}$ yields $C\left(S_{1}, S_{2} \rightarrow T_{3}\right)=1$ bit. Since $P\left(T_{1,2}\right)$ is isomorphic to $P\left(T_{1,3}\right)$, the target chain rule requires that,

$$
\begin{equation*}
C\left(S_{1}, S_{2} \rightarrow T_{1} \mid T_{3}\right)=-1 \text { bit, } \quad U\left(S_{1} \backslash S_{2} \rightarrow T_{1} \mid T_{3}\right)=1 \text { bit, } \quad U\left(S_{2} \backslash S_{1} \rightarrow T_{1} \mid T_{3}\right)=1 \text { bit. } \tag{A38}
\end{equation*}
$$

That is, one would have to accept the negative (average) PI atom $C\left(S_{1}, S_{2} \rightarrow T_{1} \mid T_{3}\right)=-1$ bit despite the fact that there are no non-zero pointwise ambiguity terms upon splitting any of $i\left(s_{1} ; t_{1} \mid t_{3}\right)$, $i\left(s_{2} ; t_{1} \mid t_{3}\right)$ and $i\left(s_{1,2} ; t_{1} \mid t_{3}\right)$ into specificity and ambiguity. Although this does not constitute a formal proof that the identity property is incompatible with the target chain rule, one would have to accept and find a way to justify $C\left(S_{1}, S_{2} \rightarrow T_{1} \mid T_{3}\right)=-1$ bit. Since there is no ambiguity in $i\left(s_{1} ; t_{1} \mid t_{3}\right), i\left(s_{2} ; t_{1} \mid t_{3}\right)$ and $i\left(s_{1,2} ; t_{1} \mid t_{3}\right)$, this result is not reconcilable within the framework of specificity and ambiguity.

## Appendix C. Additional Example Probability Distributions

## Appendix C.1. Probability Distribution Tbep

Figure A1 shows the probability distribution three bit-even parity (TBEP) which considers binary predictors variables $S_{1}, S_{2}$ and $S_{3}$ which are constrained such that together their parity is even. The target variable $T$ is simply a copy of the predictors, i.e., $T=T_{1,2,3}=\left(T_{1}, T_{2}, T_{3}\right)$ where $T_{1}=S_{1}, T_{2}=S_{2}$ and $T_{3}=S_{3}$. (Equivalently, the target can be represented by any four state variable T.) It was introduced by Bertschinger et al. [5] and revisited by Rauh et al. [13] who (as mentioned in Section 5.5) used it to prove the following by counter-example: there is no measure of redundant average information for more than two predictor variables which simultaneously satisfies the Williams and Beer Axioms, the identity property, and local positivity. The measures $I_{\mathrm{red}}, \widetilde{U I}$ and $\mathcal{S}_{\mathrm{VK}}$ these properties. Hence, this probability distribution which has been used to demonstrate that these measures are not consistent with the PID framework in the general case of an arbitrary number of predictor variables.

This example is similar to TBC in the several ways. Firstly, due to the symmetry in the probability distribution, each realisation will have the same pointwise decomposition. Secondly, there is an isomorphism between the probability distributions $P(T)$ and $P\left(S_{1}, S_{2}, S_{3}\right)$, and hence the pointwise ambiguity provided by any (individual or joint) predictor event is 0 bit (since given $t$, one knows $s_{1}, s_{2}$ and $s_{3}$ ). Thirdly, the individual predictor events $s_{1}, s_{2}$ and $s_{3}$ each exclude $1 / 2$ of the total probability mass in $P(T)$ and so each provide 1 bit of pointwise specificity. Thus, there is 1 bit of three-way redundant, pointwise specificity in each realisation. Fourthly, the joint predictor event $s_{1,2,3}$ excludes $3 / 4$ of the total probability mass, providing 2 bit of pointwise specificity (which is similar to TBC). However, unlike TBC, one could consider the three joint predictor events $s_{1,2}, s_{1,3}$ and $s_{2,3}$. These joint pairs also exclude $3 / 4$ of the total probability mass each, and hence also each provide 2 bit of pointwise specificity. As such, there is 1 bit of pointwise, three-way redundant, pairwise complementary specificity between these three joint pairs of source events, in addition to the 1 bit of three-way redundant, pointwise specificity. Finally, putting this together and averaging over all realisations, TBEP consists of 1 bit of three-way redundant information and 1 bit of three-way redundant, pairwise complementary information. The resultant average decomposition is the same as the decomposition induced by $I_{\min }$ [5].

| $P\left(S_{1,2,3}, T\right)$ |  |  |
| :--- | :---: | :---: |
| 0 | $1 / 4$ | 000 |
| 1 | $1 / 4$ | 011 |
| 2 | $1 / 4$ | 101 |
| 3 | $1 / 4$ | 110 |






Figure A1. Example Tber. (Top) probability mass diagram for realisation ( $S_{1}=0, S_{2}=0, S_{3}=0, T=000$ ); (Bottom left) With three predictors, it is convenient to represent to decomposition diagrammatically. This is especially true TBEP as one only needs to consider the specificity lattice for one realisation; (Bottom right) The specificity lattice for the realisation ( $S_{1}=0, S_{2}=0, S_{3}=0, T=000$ ). For each source event the left value corresponds to the value of $i_{\cap}^{+}$, evaluated using $r_{\min ^{+}}^{+}$, while the right value (surrounded by parenthesis) corresponds to the partial information $\pi^{+}$.

## Appendix C.2. Probability Distribution UNQ

Figure A2 shows the decomposition of the probability distribution unique (UNQ). Note that this probability distribution corresponds to RDNERR where the error probability $\varepsilon=1 / 2$, and hence the similarity in the resultant distributions. The results may initially seem unusual, that the predictor $S_{1}$ is not uniquely informative since $U\left(S_{1} \backslash S_{2} \rightarrow T\right)=0$ bit as one might intuitively expect. Rather it is deemed to be redundantly informative $R I=1$ bit with the predictor $S_{2}$ which is also uniquely misinformative $U\left(S_{2} \backslash S_{1} \rightarrow T\right)=-1$ bit. This is because both $S_{1}$ and $S_{2}$ provide $I^{+}\left(S_{1} \rightarrow T\right)=I^{+}\left(S_{2} \rightarrow T\right)=1$ bit of specificity; however the information provided by $S_{2}$ is unique in that the 1 bit provided is not "useful" ([42], p. 21) and hence $I\left(S_{2} \rightarrow T\right)=1$ bit while $I\left(S_{2} \rightarrow T\right)=1$ bit. Finally, the complementary information $C\left(S_{1}, S_{2} \rightarrow T\right)=1$ bit is required by the decomposition in order to balance this 1 bit of unique ambiguity. The results in this example partly explain our preference for term complementary information as opposed to synergistic information-while $C\left(S_{1}, S_{2} \rightarrow T\right)=1$ bit is readily explainable, it would be dubious to refer to this as synergy given that $S_{1}$ enables perfect predictions of $T$ without any knowledge of $S_{2}$.


| $p$ | $s_{1}$ | $s_{2}$ | $t$ | $i_{1}^{+}$ | $i_{1}^{-}$ | $i_{2}^{+}$ | $i_{2}^{-}$ | $i_{12}^{+}$ | $i_{12}^{-}$ | $r^{+}$ | $u_{1}^{+}$ | $u_{2}^{+}$ | $c^{+}$ | $r^{-}$ | $u_{1}^{-}$ | $u_{2}^{-}$ | $c^{-}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1/4 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 1/4 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 1/4 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 1/4 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| Expected values |  |  |  | 1 | 0 | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |

$R\left(S_{1}, S_{2} \rightarrow T\right)=1$ bit $\quad U\left(S_{1} \backslash S_{2} \rightarrow T\right)=0$ bit $\quad U\left(S_{2} \backslash S_{1} \rightarrow T\right)=-1$ bit $\quad C\left(S_{1}, S_{2} \rightarrow T\right)=1$ bit

Figure A2. Example UnQ. (Top) the probability mass diagrams for every single possible realisation; (Middle) for each realisation, the PPID using specificity and ambiguity is evaluated (see Figure 4); (Bottom) the atoms of (average) partial infromation obtained through recombination of the averages.

## Appendix C.3. Probability Distribution AND

Figure A3 shows the decomposition of the probability distribution and (AND). Note that the probability distribution or (OR) has the same decomposition as the target distributions are isomorphic.


| $p$ | $s_{1}$ | $s_{2}$ | $t$ | $i_{1}^{+}$ | $i_{1}^{-}$ | $i_{2}^{+}$ | $i_{2}^{-}$ | $i_{12}^{+}$ | $i_{12}^{-}$ | $r^{+}$ | $u_{1}^{+}$ | $u_{2}^{+}$ | $c^{+}$ | $r^{-}$ | $u_{1}^{-}$ | $u_{2}$ | $c^{-}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1/4 | 0 | 0 | 0 | 1 | $\lg 3 / 2$ | 1 | $\lg 3 / 2$ | 2 | $\lg 3$ | 1 | 0 | 0 | 1 | $\lg 3 / 2$ | 0 | 0 | 1 |
| 1/4 | 0 | 1 | 0 | 1 | $\lg 3 / 2$ | 1 | $\lg 3$ | 2 | $\lg 3$ | 1 | 0 | 0 | 1 | $\lg 3 / 2$ | 0 | 1 | 0 |
| 1/4 | 1 | 0 | 0 | 1 | $\lg 3$ | 1 | $\lg 3 / 2$ | 2 | $\lg 3$ | 1 | 0 | 0 | 1 | $\lg 3 / 2$ | 1 | 0 | 0 |
| 1/4 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Expected values |  |  |  | 1 | 0.689 | 1 | 0.689 | 2 | 1.189 | 1 | 0 | 0 | 1 | 0.439 | 0.250 | 0.250 | 0.25 |

$R\left(S_{1}, S_{2} \rightarrow T\right)=0.561$ bit $\quad U\left(S_{1} \backslash S_{2} \rightarrow T\right)=-0.25$ bit $U\left(S_{2} \backslash S_{1} \rightarrow T\right)=-0.25$ bit $\quad C\left(S_{1}, S_{2} \rightarrow T\right)=0.75$ bit
Figure A3. Example AND. (Top) the probability mass diagrams for every single possible realisation; (Middle) for each realisation, the PPID using specificity and ambiguity is evaluated (see Figure 4); (Bottom) the atoms of (average) partial infromation obtained through recombination of the averages.

## References and Note

1. Williams, P.L.; Beer, R.D.Information decomposition and synergy. Nonnegative decomposition of multivariate information. arXiv 2010, arXiv:1004.2515.
2. Williams, P.L.; Beer, R.D. Indiana University. Decomposing Multivariate Information. Privately communicated, 2010. This unpublished paper is highly similar to [1]. Crucially, however, this paper derives the redundancy lattice from the W\&B Axioms 1-3 of Section 1. In contrast, [1] derives the redundancy lattice as a property of the particular measure $I_{\text {min }}$.
3. Olbrich, E.; Bertschinger, N.; Rauh, J. Information decomposition and synergy. Entropy 2015, 17, 3501-3517. [CrossRef]
4. Lizier, J.T.; Flecker, B.; Williams, P.L. Towards a synergy-based approach to measuring information modification. In Proceedings of the IEEE Symposium on Artificial Life (ALife), Singapore, 16-19 April 2013; pp. 43-51.
5. 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, Brussels, Belgium, 3-7 September 2012; Springer: Cham, The Netherland, 2013; pp. 251-269.
6. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. Phys. Rev. E 2013, 87, 012130. [CrossRef]
7. Griffith, V.; Chong, E.K.; James, R.G.; Ellison, C.J.; Crutchfield, J.P. Intersection information based on common randomness. Entropy 2014, 16, 1985-2000. [CrossRef]
8. Shannon, C.E. A Mathematical Theory of Communication. Bell Syst. Tech. J. 1948, 27, 379-423. [CrossRef]
9. Fano, R. Transmission of Information; The MIT Press: Cambridge, MA, USA, 1961.
10. Harder, M. Information driven self-organization of agents and agent collectives. Ph.D. Thesis, University of Hertfordshire, Hertfordshire, UK, 2013.
11. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. Entropy 2014, 16, 2161-2183. [CrossRef]
12. Griffith, V.; Koch, C. Quantifying Synergistic Mutual Information. In Guided Self-Organization: Inception; Prokopenko, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 9, pp. 159-190.
13. 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, Honolulu, HI, USA, 29 June-4 July 2014; pp. 2232-2236.
14. Perrone, P.; Ay, N. Hierarchical Quantification of Synergy in Channels. Front. Robot. AI 2016, 2, 35. [CrossRef]
15. Griffith, V.; Ho, T. Quantifying redundant information in predicting a target random variable. Entropy 2015, 17, 4644-4653. [CrossRef]
16. Rosas, F.; Ntranos, V.; Ellison, C.J.; Pollin, S.; Verhelst, M. Understanding interdependency through complex information sharing. Entropy 2016, 18, 38. [CrossRef]
17. Barrett, A.B. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. Phys. Rev. E 2015, 91, 052802. [CrossRef]
18. Ince, R. Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. Entropy 2017, 19, 318. [CrossRef]
19. Ince, R.A. The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. arXiv 2017, arXiv:1702.01591.
20. Chicharro, D.; Panzeri, S. Synergy and Redundancy in Dual Decompositions of Mutual Information Gain and Information Loss. Entropy 2017, 19, 71. [CrossRef]
21. Rauh, J.; Banerjee, P.K.; Olbrich, E.; Jost, J.; Bertschinger, N. On Extractable Shared Information. Entropy 2017, 19, 328. [CrossRef]
22. Rauh, J.; Banerjee, P.K.; Olbrich, E.; Jost, J.; Bertschinger, N.; Wolpert, D. Coarse-Graining and the Blackwell Order. Entropy 2017, 19, 527. [CrossRef]
23. Rauh, J. Secret sharing and shared information. Entropy 2017, 19, 601. [CrossRef]
24. Faes, L.; Marinazzo, D.; Stramaglia, S. Multiscale information decomposition: Exact computation for multivariate Gaussian processes. Entropy 2017, 19, 408. [CrossRef]
25. Pica, G.; Piasini, E.; Chicharro, D.; Panzeri, S. Invariant components of synergy, redundancy, and unique information among three variables. Entropy 2017, 19, 451. [CrossRef]
26. James, R.G.; Crutchfield, J.P. Multivariate dependence beyond shannon information. Entropy 2017, 19, 531. [CrossRef]
27. Makkeh, A.; Theis, D.O.; Vicente, R. Bivariate Partial Information Decomposition: The Optimization Perspective. Entropy 2017, 19, 530. [CrossRef]
28. Kay, J.W.; Ince, R.A.; Dering, B.; Phillips, W.A. Partial and Entropic Information Decompositions of a Neuronal Modulatory Interaction. Entropy 2017, 19, 560. [CrossRef]
29. Angelini, L.; de Tommaso, M.; Marinazzo, D.; Nitti, L.; Pellicoro, M.; Stramaglia, S. Redundant variables and Granger causality. Phys. Rev. E 2010, 81, 037201. [CrossRef]
30. Stramaglia, S.; Angelini, L.; Wu, G.; Cortes, J.M.; Faes, L.; Marinazzo, D. Synergetic and redundant information flow detected by unnormalized Granger causality: Application to resting state fMRI. IEEE Trans. Biomed. Eng. 2016, 63, 2518-2524. [CrossRef]
31. Ghazi-Zahedi, K.; Langer, C.; Ay, N. Morphological computation: Synergy of body and brain. Entropy 2017, 19, 456. [CrossRef]
32. Maity, A.K.; Chaudhury, P.; Banik, S.K. Information theoretical study of cross-talk mediated signal transduction in MAPK pathways. Entropy 2017, 19, 469. [CrossRef]
33. Tax, T.; Mediano, P.A.; Shanahan, M. The partial information decomposition of generative neural network models. Entropy 2017, 19, 474. [CrossRef]
34. Wibral, M.; Finn, C.; Wollstadt, P.; Lizier, J.T.; Priesemann, V. Quantifying Information Modification in Developing Neural Networks via Partial Information Decomposition. Entropy 2017, 19, 494. [CrossRef]
35. Woodward, P.M. Probability and Information Theory: With Applications to Radar; Pergamon Press: Oxford, UK, 1953.
36. Woodward, P.M.; Davies, I.L. Information theory and inverse probability in telecommunication. Proc. IEE-Part III Radio Commun. Eng. 1952, 99, 37-44. [CrossRef]
37. Gray, R.M. Probability, Random Processes, and Ergodic Properties; Springer: New York, NY, USA, 1988.
38. Martin, N.F.; England, J.W. Mathematical Theory of Entropy; Cambridge University Press: Cambridge, UK, 1984.
39. Finn, C.; Lizier, J.T. Probability Mass Exclusions and the Directed Components of Pointwise Mutual Information. arXiv 2018, arXiv:1801.09223.
40. Kelly, J.L. A new interpretation of information rate. Bell Labs Tech. J. 1956, 35, 917-926. [CrossRef]
41. Ash, R. Information Theory; Interscience tracts in pure and applied mathematics; Interscience Publishers: Geneva, Switzerland, 1965.
42. Shannon, C.E.; Weaver, W. The Mathematical Theory of Communication; University of Illinois Press: Champaign, IL, USA, 1998.
43. Cover, T.M.; Thomas, J.A. Elements of Information Theory; John Wiley \& Sons: Hoboken, NJ, USA, 2012.
44. Pearl, J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1988.
45. Rota, G.C. On the foundations of combinatorial theory I. Theory of Möbius functions. Probab. Theory Relat. Field 1964, 2, 340-368.
46. Stanley, R.P. Enumerative Combinatorics. In Cambridge Studies in Advanced Mathematics, 2nd ed.; Cambridge University Press: Cambridge, UK, 2012; Volume 1.
47. Davey, B.A.; Priestley, H.A. Introduction to Lattices and Order, 2nd ed.; Cambridge University Press: Cambridge, UK, 2002.
48. Ross, S.M. A First Course in Probability, 8th ed.; Pearson Prentice Hall: Upper Saddle River, NJ, USA, 2009.

## Erratum

1. The inequality in W\&B Axiom 2 on page 3 of the published work is the wrong way around.
2. The inequality in PPID Axiom 2 on page 7 of the published work is the wrong way around.
3. The partial specificity and partial ambiguity are respectively denoted $\pi^{+}$and $\pi^{-}$in some places, and $1_{\partial}^{+}$and $i_{\partial}^{-}$in others. This issue affects Theorems 3, 4 and A2 and their respective proofs, Equations (A16-A26), and Figure A1.
4. The Equation (A16) should read

$$
\begin{equation*}
i_{\cap}^{ \pm}(\alpha \rightarrow t)=\sum_{\alpha \leq \beta} i_{\partial}^{ \pm}(\alpha \rightarrow t) . \tag{E1}
\end{equation*}
$$

## Chapter 5

## A New Framework for Information Decomposition

The paper from Chapter 4 constructs its decomposition from the top-down-that is to say, the a set of axioms is specified, a function satisfying these axioms is given, and its various properties are derived. On the other hand, the paper we present in this chapter provides a bottom-up derivation of what turns out to be an equivalent decomposition. We begin by considering the following idea: suppose that two individuals, Alice and Bob, each respectively observe the marginal realisations $x$ and $y$, and say that they then share their information content with a third non-observing party. This observer, Eve, knows which joint realisation has occurred and she knows the marginal probability distributions, but she does not know the joint distribution. We then ask how much information does Eve have?

It turns out that Eve's information must be given by the maximum of the Alice's and Bob's information. However, this is deceptively simple. Specifically, as we consider including more marginal observers (i.e. Alice- and Bob-like observers), who are each sharing their information with nonobservers (i.e. Eve-like individuals), we find an entire family of new measures of shared marginal information that possess a non-trivial algebraic structure. This structure is a distributive lattice that, as was discussed in Chapter 2, possess both an intersection- and union-like operators which are idempotent, commutative, associative, and distributive, and are connected by absorption. Each distinct way that a set of Alice- and Bob-like marginal observers can share their information with an Eve-like individual corresponds to a new measure of shared marginal information content, which in turn corresponds to an element in a distributive lattice. Moreover, as a consequence of the fundamental theorem of distributive lattices, this lattice of shared marginal information content is isomorphic to the set union and intersections. This is the key result from this chapter, as it means that these new measures of information content can be represented precisely with Venn diagrams. As discussed in Chapter 2, this was a significant issue for the mutual and multivariate mutual information content, as they cannot be accurately depicted for more than two variables.

Building upon this result, we then combine the structure of joint information content together with the newly-introduced distributive lattice of shared marginal information content to form one overall algebraic structure. This structure is highly non-trivial, and is not explored in detail. However, we do show that the redundancy lattice fro $m$ partial information decomposition is embedded within this larger algebraic structure. To be specific, since we are considering marginal information contents, this structure is actually equivalent to the specificity lattice from Chapter 4 . The chapter closes by discussing whether or not one should combine the information contents from the specificity and ambiguity lattices.

## Article

# Generalised Measures of Multivariate Information Content 

Conor Finn ${ }^{1,2, *(\mathbb{D}}$ and Joseph T. Lizier ${ }^{1}$ (D)<br>1 Centre for Complex Systems, The University of Sydney, Sydney NSW 2006, Australia; joseph.lizier@sydney.edu.au<br>2 CSIRO Data61, Marsfield NSW 2122, Australia<br>* Correspondence: conor.finn@sydney.edu.au

Received: 11 December 2019; Accepted: 12 February 2020; Published: 14 February 2020


#### Abstract

The entropy of a pair of random variables is commonly depicted using a Venn diagram. This representation is potentially misleading, however, since the multivariate mutual information can be negative. This paper presents new measures of multivariate information content that can be accurately depicted using Venn diagrams for any number of random variables. These measures complement the existing measures of multivariate mutual information and are constructed by considering the algebraic structure of information sharing. It is shown that the distinct ways in which a set of marginal observers can share their information with a non-observing third party corresponds to the elements of a free distributive lattice. The redundancy lattice from partial information decomposition is then subsequently and independently derived by combining the algebraic structures of joint and shared information content.


Keywords: information content; multivariate mutual information; information measures; information decomposition; synergy; redundancy

## 1. Introduction

For any pair of random variables $X$ and $Y$, the entropy $H$ satisfies the inequality

$$
\begin{equation*}
H(X)+H(Y) \geq H(X, Y) \geq H(X), H(Y) \geq 0 \tag{1}
\end{equation*}
$$

From this inequality, it is easy to see that the conditional entropies and mutual information are non-negative,

$$
\begin{align*}
H(X \mid Y) & =H(X, Y)-H(Y) \geq 0  \tag{2}\\
H(Y \mid X) & =H(X, Y)-H(X) \geq 0  \tag{3}\\
I(X ; Y) & =H(X)+H(Y)-H(X, Y) \geq 0 \tag{4}
\end{align*}
$$

For any pair of sets $A$ and $B$, a measure $\mu$ satisfies the inequality

$$
\begin{equation*}
\mu(A)+\mu(B) \geq \mu(A \cup B) \geq \mu(A), \mu(B) \geq 0 \tag{5}
\end{equation*}
$$

which follows from the non-negativity of measure on the relative complements and the intersection,

$$
\begin{align*}
& \mu(A \backslash B)=\mu(A \cup B)-\mu(B) \geq 0  \tag{6}\\
& \mu(B \backslash A)=\mu(A \cup B)-\mu(A) \geq 0  \tag{7}\\
& \mu(A \cap B)=\mu(A)+\mu(B)-\mu(A \cup B) \geq 0 . \tag{8}
\end{align*}
$$

Although the entropy is not itself a measure, several authors have noted the entropy is analogous to measure in this regard [1-7]. Indeed, it is this analogy which provides the justification for the typical depiction of a pair of entropies using Venn diagrams, i.e., Figure 1. Nevertheless, MacKay [8] noted that this representation is misleading for at least two reasons: Firstly, since the measure on the intersection $\mu(A \cap B)$ is a measure on a set, it gives the false impression that the mutual information $I(X ; Y)$ is the entropy of some intersection between the random variables. Secondly, it might lead one to believe that this analogy can be generalised beyond two variables. However, the analogy does not generalise beyond two variables since the multivariate mutual information [9] between three random variables (which is also known as the interaction information [10], amount of information [2] or co-information [11]),

$$
\begin{equation*}
I(X ; Y ; Z)=H(X)+H(Y)+H(Z)-H(X, Y)-H(X, Z)-H(Y, Z)+H(Z, Y, Z) \tag{9}
\end{equation*}
$$

is not non-negative $[3,9,12]$, and hence is not analogous to measure on the triple intersection $\mu(A \cap$ $B \cap C)$ [3]. Indeed, this "unfortunate" property led Cover and Thomas to conclude that "there isn't really a notion of mutual information common to three random variables" (p. 49 [13]). Consequently, MacKay [8] recommended against depicting the entropy of three or more variables using a Venn diagram, i.e., Figure 1, unless one is aware of these issues with this representation.


Figure 1. (Top left) When depicting a measure on the union of two sets $\mu(A \cup B)$, the area of each section can be used to represent the inequality (5) and hence the values $\mu(A \backslash B), \mu(B \backslash A)$ and $\mu(A \cap B)$ correspond to the area of each section. This correspondence can be generalised to consider an arbitrary number of sets. (Bottom left) When depicting the joint entropy $H(X, Y)$, the area of each section can also be used to represent the inequality (1) and hence the values $H(X \mid Y), H(Y \mid X)$ and $I(X ; Y)$ correspond to the area of each section. However, this correspondence does not generalise beyond two variables. (Right) For example, when considering the entropy of three variables, the multivariate mutual information $I(X ; Y ; Z)$ cannot be accurately represented using an area since, as represented by the hatching, it is not non-negative.

However, Yeung [6,7] showed that there is an analogy between entropy and signed measure that is valid for an arbitrary number of random variables. To do this, Yeung defined a signed measure on a suitably constructed $\sigma$-algebra that is uniquely determined by the joint entropies of the random variables involved. This correspondence enables one to establish information-theoretic identities from measure-theoretic identities and hence Venn diagrams can be used to represent the entropy of three or
more variables provided one is aware that the certain overlapping areas may correspond to negative quantities. Moreover, the multivariate mutual information is useful both as summary quantity and for manipulating information-theoretic identities provided one is mindful it may have "no intuitive meaning" $[5,6]$.

In this paper, we introduce new measures of multivariate information that are analogous to measures upon sets and maintain their operational meaning when considering an arbitrary number of variables. These new measures complement the existing measures of multivariate mutual information, and will be constructed by considering the distinct ways in which a set of marginal observers might share their information with a non-observing third party. In Section 2, we discuss the existing measures of information content in terms of a set of individuals who each have different knowledge about a joint realisation from a pair of random variables. Then, in Section 3, we discuss how these individuals can share their information with a non-observing third party, and derive the functional form of this individual's information. In Section 4, we relate this new measure of information content back to the mutual information. Sections 5-7 then generalise the arguments of Sections 3 and 4 to consider an arbitrary number of observers. Finally, in Section 8, we discuss how these new measures can be combined to define new measures of mutual information.

## 2. Mutual Information Content

Suppose that Alice and Bob are separately observing some process and let the discrete random variables $X$ and $Y$ represent their respective observations. Say that Johnny is a third individual who can simultaneously make the same observations as Alice and Bob such that his observations are given by the joint variable $(X, Y)$. When a realisation $(x, y)$ occurs, Alice's information is given by the information content [8],

$$
\begin{equation*}
h(x)=-\log p_{X}(x) \geq 0 \tag{10}
\end{equation*}
$$

where $p_{X}(x)$ is the probability mass of the realisation $x$ of variable $X$ computed from the probability distribution $p_{X}$. Likewise, Bob's information is given by the information content $h(y)$, while Johnny's information is by the joint information content $h(x, y)=-\log p_{X Y}(x, y)$. The information that Alice can expect to gain from an observation is given by the entropy,

$$
\begin{equation*}
H(X)=\mathrm{E}_{X}[h(x)] \geq 0 \tag{11}
\end{equation*}
$$

where $E_{X}$ represents an expectation value over realisations of the variable $X$. Similarly, Bob's expected information gain is given by the entropy $H(X)$ and Johnny's expected information is given by the joint entropy $H(X, Y)=\mathrm{E}_{X Y}[h(x, y)]$. Clearly, for any realisation, Johnny has at least as much information as either Alice or Bob,

$$
\begin{equation*}
h(x, y) \geq h(x), h(y) \geq 0 \tag{12}
\end{equation*}
$$

The conditional information content can be used to quantify how much more information Johnny has relative to either Alice or Bob, respectively,

$$
\begin{align*}
& h(x \mid y)=h(x, y)-h(y) \geq 0  \tag{13}\\
& h(y \mid x)=h(x, y)-h(x) \geq 0 \tag{14}
\end{align*}
$$

Similarly, we can quantify how much more information Johnny expects to get compared to either Alice or Bob via the conditional entropies,

$$
\begin{align*}
& H(X \mid Y)=\mathrm{E}_{X Y}[h(x \mid y)] \geq 0  \tag{15}\\
& H(Y \mid X)=\mathrm{E}_{X Y}[h(y \mid x)] \geq 0 \tag{16}
\end{align*}
$$

Now, consider a fourth individual who does not directly observe the process, but with whom Alice and Bob share their knowledge. To be explicit, we are considering the situation whereby this
individual knows that the joint realisation $(x, y)$ has occurred and knows the marginal distributions $p_{X}$ and $p_{Y}$, but does not know the joint distribution $p_{X Y}$. How much information does this individual obtain from the shared marginal knowledge provided by Alice and Bob? The answer to this question is provided in Section 3, but for now let us consider a simplified version of this problem. Suppose that such an individual, whom we call Indiana (or Indy for short), assumes that Alice's observations are independent of Bob's observations. In terms of the probabilities, this means that Indy believes that the joint probability $p_{X Y}(x, y)$ is equal to the product probability $p_{X \times Y}(x, y)=p_{X}(x) p_{Y}(y)$, while, in terms of information, this assumption leads Indiana to believe that her information is given by the independent information content $h(x)+h(y)$. Moreover, the information that Indiana expects to gain from any one realisation is given by $H(X)+H(Y)$.

Let us now compare how much information Indiana believes that she has compared to our other observers. For every realisation, Indiana believes that she has at least as much information as either Alice or Bob,

$$
\begin{equation*}
h(x)+h(y) \geq h(x), h(y) \geq 0 \tag{17}
\end{equation*}
$$

Since Indy knows what both Alice and Bob know individually, it is hardly surprising that she always has at least as much information as either Alice or Bob. The comparison between Indiana and Johnny, however, is not so straightforward-there is no inequality that requires the information content of the joint realisation to be less than the information content of the independent realisations, or vice versa. Consequently, the difference between the information that Indiana thinks she has and Johnny's information, i.e., the mutual information content between a pair of realisations,

$$
\begin{equation*}
i(x ; y)=h(x)+h(y)-h(x, y)=\log \frac{p_{X Y}(x, y)}{p_{X}(x) p_{Y}(y)} \tag{18}
\end{equation*}
$$

is not non-negative [14]. (This function goes by several different names including the pointwise mutual information, the information density [15] or simply the mutual information [9].) Thus, similar to how it is potentially misleading to depict the entropy of three of more variables using a Venn diagram, representing the information content of two variables using a Venn diagram is somewhat dubious (see Figure 2).


Figure 2. (Left) Indiana assumes that Alice's information $h(x)$ is independent of Bob's information $h(y)$ such that her information is given by $h(x)+h(y)$. (Middle) Johnny knows the joint distribution $p_{X Y}$, and hence his information is given by the joint information content $h(x, y)$. (Right) There is no inequality that requires Johnny's information to be no greater than Indiana's assumed information, or vice versa. On the one hand, Johnny can have more information than Indiana since a joint realisation can be more surprising than both of the individual marginal realisations. On the other hand, Indiana can have more information than Johnny since a joint realisation can be less surprising than both of the individual marginal realisations occurring independently. Thus, as represented by the hatching, the mutual information content $i(x ; y)$ is not non-negative.

Since Johnny knows the joint distribution $p_{X Y}$, while Indiana only knows the marginal distributions $p_{X}(x)$ and $p_{Y}(y)$, we might expect that Indiana should never have more information than Johnny. However, Indiana's assumed information is based upon the belief that Alice's observations $X$ are independent of Bob's observations $Y$, which leads Indiana to overestimate her information on average.

Indeed, Indiana is so optimistic that the information she expects to get upper bounds the information that Johnny can expect to get,

$$
\begin{equation*}
H(X)+H(Y) \geq H(X, Y) \geq 0 \tag{19}
\end{equation*}
$$

Thus, despite the fact that Indiana can have less information than Johnny for certain realisations-i.e., despite the fact that the mutual information content is not non-negative-the mutual information in expectation is non-negative,

$$
\begin{equation*}
I(X ; Y)=H(X)+H(Y)-H(X, Y)=\mathrm{E}_{X Y}[i(x ; y)] \geq 0 \tag{20}
\end{equation*}
$$

Crucially, and in contrast to the information content (10) and entropy (11), the non-negativity of the mutual information does not follow directly from the non-negativity of the mutual information content (18), but rather must be proved separately. (Typically, this is done by showing that the mutual information can be written as a Kullback-Leibler divergence which is non-negative by Jensen's inequality, e.g., see Cover and Thomas [13].) Thus, not only does Indiana potentially have more information than Johnny for certain realisations, but on average we expect Indiana to have more information than Johnny. Of course, by assuming Alice's observations are independent of Bob's observations, Indiana is overestimating her information. Thus, in the next section, we consider the situation whereby one does not make this assumption.

## 3. Marginal Information Sharing

Suppose that Eve is another individual who, similar to Indiana, does not make any direct observations, but with whom both Alice and Bob share their knowledge; i.e., Eve knows the joint realisation $(x, y)$ has occurred and knows the marginal distributions $p_{X}$ and $p_{Y}$, but does not know the joint distribution $p_{X Y}$. Furthermore, suppose that Eve is more conservative than Indiana and does not assume that Alice's observations are independent of Bob's observations-how much information does Eve have for any one realisation?

It seems clear that Eve's information should always satisfy the following two requirements. Firstly, since Alice and Bob both share their knowledge with Eve, she should have at least as much information as either of them have individually. Secondly, since Eve has less knowledge than Johnny, she should have no more information than Johnny; i.e., in contrast to Indy, Eve should never have more information than Johnny. As the following theorem shows, these two requirements uniquely determine the functional form of Eve's information:

Theorem 1. The unique function $h(x \sqcup y)$ of $p_{X}(x)$ and $p_{Y}(y)$ that satisfies $h(x, y) \geq h(x \sqcup y) \geq$ $h(x), h(y) \geq 0$ for all $p_{X Y}(x, y)$ is

$$
\begin{equation*}
h(x \sqcup y)=\max (h(x), h(y)) \geq 0 \tag{21}
\end{equation*}
$$

Proof. Clearly, the function is lower bounded by max $(h(x), h(y))$. The upper bound is given by the minimum possible $h(x, y)$, which corresponds to the maximum allowed $p_{X Y}(x, y)$. For any $p_{X}(x)$ and $p_{Y}(y)$, the maximum allowed $p_{X Y}(x, y)$ is $\min \left(p_{X}(x), p_{Y}(y)\right)$, which corresponds to $h(x, y)=$ $\max (h(x), h(y))$.

Eve's information is given by the maximum of Alice's and Bob's information, or the information content of the most surprising marginal realisation. Although we have defined Eve's information by requiring it to be no greater than Johnny's information, it is also clear that Eve also has no more information than Indiana. As such, Eve's information satisfies the inequality

$$
\begin{equation*}
h(x)+h(y) \geq h(x \sqcup y) \geq h(x), h(y) \geq 0 \tag{22}
\end{equation*}
$$

which is analogous to the inequality (5) satisfied by measure. Hence, as pre-empted by the notation (and as further justified in Section 6), Eve's information is referred to as the union information content. The union information content is the maximum possible information that Eve can get from knowing what Alice and Bob know-it quantifies the information provided by a joint event $(x, y)$ when one knows the marginal distributions $p_{X}$ and $p_{Y}$, but does not know nor make any assumptions about the joint distribution $p_{X Y}$.

Similar to how the conditional information contents (15) and (16) enable us to quantify how much more information Johnny has relative to either Alice or Bob, the inequality (22) enables us to quantify how much information Eve gets from Alice relative to Bob and vice versa, respectively,

$$
\begin{align*}
& h(x \backslash y)=h(x \sqcup y)-h(y)=\max (h(x)-h(y), 0) \geq 0,  \tag{23}\\
& h(y \backslash x)=h(x \sqcup y)-h(x)=\max (0, h(y)-h(x)) \geq 0 . \tag{24}
\end{align*}
$$

These non-negative functions are analogous to measure on the relative complements of a pair of sets and are called the unique information content from $x$ relative to $y$, and vice versa, respectively. It is easy to see that, since Eve's information is either equal to Alice's or Bob's information (or both), at least one of these two functions must be equal to zero.

The inequality (22) also enables us to quantify how much more information Indiana has relative to Eve. Since Indiana's assumed information is given by the sum of Alice's and Bob's information while Eve's information is given by the maximum of Alice's and Bob's information, the difference between the two is given by the minimum of Alice's and Bob's information,

$$
\begin{equation*}
h(x \sqcap y)=h(x)+h(y)-h(x \sqcup y)=h(x)+h(y)-\max (h(x), h(y))=\min (h(x), h(y)) \geq 0 \tag{25}
\end{equation*}
$$

In contrast to the comparison between Indiana and Johnny, i.e., the mutual information content (18), the comparison between Indiana and Eve is non-negative. As such, this function is analogous to measure on the intersection of two sets and hence will be referred to as the intersection information content. The intersection information content is the minimum possible information that Eve could have gotten from knowing either what Alice or Bob know, and is given by the information content of the least surprising marginal realisation.

Finally, from (21) and (23)-(25), it is not difficult to see that Eve's information can be decomposed into the information that could have been obtained from either Alice or Bob, the unique information from Alice relative to Bob and the unique information from Bob relative to Alice,

$$
\begin{equation*}
h(x \sqcup y)=h(x \sqcap y)+h(x \backslash y)+h(y \backslash x) . \tag{26}
\end{equation*}
$$

Of course, as already discussed, at least one of these unique information contents must be zero. Figure 3 depicts this decomposition for some realisation whereby Alice's information $h(x)$ is greater than Bob's information $h(y)$.

To summarise thus far, both Alice and Bob share their information with Indiana and Eve, who then each interpret this information in a different way. By comparing Figures 2 and 3, we can easily contrast their distinct perspectives. Eve is more conservative than Indiana and assumes that she has gotten as little information as she could possibly have gotten from knowing what Alice and Bob know; this is given by the maximum from Alice's and Bob's information, or is the information content associated with the most surprising marginal realisation observed by Alice and Bob. In effect, Eve's conservative approach means that she pessimistically assumes that the information provided by the least surprising marginal realisation was already provided by the most surprising marginal realisation. In contrast, Indiana optimistically assumes that the information provided by the least surprising marginal realisation is independent of the information provided by the most surprising marginal realisation.


Figure 3. (Left) If Alice's information $h(x)$ is greater than Bob's information $h(y)$, then Eve's information $h(x \sqcup y)$ is equal to Alice's information $h(x)$. In effect, Eve is pessimistically assuming that information provided by the least surprising marginal realisation $h(x \sqcap y)$ is already provided by the most surprising marginal realisation $h(x \sqcup y)$, i.e., Bob's information $h(y)$ is a subset of Alice's information $h(x)$. From this perspective, Eve gets unique information from Alice relative to Bob $h(x \backslash y)$, but does not get any unique information from Bob relative to Alice $h(y \backslash x)=0$. (Right) Although for each realisation Eve can only get unique information from either Alice or Bob, it is possible that Eve can expect to get unique information from both Alice and Bob on average. (Do not confuse this representation of the union entropy with the diagram that represents the joint entropy in Figure 1).

Let us now consider the information that Eve expects to get from a single realisation,

$$
\begin{equation*}
H(X \sqcup Y)=\mathrm{E}_{X Y}[h(x \sqcup y)] \geq 0 \tag{27}
\end{equation*}
$$

This function is called the union entropy, and quantifies the expected surprise of the most surprising realisation from either $X$ or $Y$. Similar to how the non-negativity of the entropy (11) follows from the non-negativity of the information content (10), the non-negativity of the union entropy (27) follows directly from the non-negativity of the union information content (21)-i.e., we do not need to invoke Jensen's inequality. Indeed, the union entropy cannot be written as a Kullback-Leibler divergence.

Since the expectation value is monotonic, and since the union information content satisfies the inequality (22), we get that the union entropy satisfies

$$
\begin{equation*}
H(X)+H(Y) \geq H(X \sqcup Y) \geq H(X), H(Y) \geq 0 \tag{28}
\end{equation*}
$$

and hence is also analogous to measure on the union of two sets. Using this inequality, we can quantify how much more information Eve expects to get from Alice relative to Bob, or vice versa, respectively,

$$
\begin{align*}
& H(X \backslash Y)=H(X \sqcup Y)-H(Y)=\mathrm{E}_{X Y}[h(x \backslash y)] \geq 0,  \tag{29}\\
& H(Y \backslash X)=H(X \sqcup Y)-H(X)=\mathrm{E}_{X Y}[h(y \backslash x)] \geq 0, \tag{30}
\end{align*}
$$

These functions are also analogous to measure on the relative complements of a pair of sets and hence will be called the unique entropy from $X$ relative to $Y$, and vice versa, respectively. Crucially, and in contrast to (23) and (24), both of these quantities can be simultaneously non-zero; although Alice might observe the most surprising event in one joint realisation, Bob might observe the most surprising event in another and hence both functions can be simultaneously non-zero.

Now, consider how much more information Indiana expects to get relative to Eve,

$$
\begin{equation*}
H(X \sqcap Y)=H(X)+H(Y)-H(X \sqcup Y)=\mathrm{E}_{X Y}[h(x \sqcap y)] \geq 0 \tag{31}
\end{equation*}
$$

This function is also analogous to measure on the intersection of two sets function will be called the intersection entropy. In contrast to the mutual information (20), since the intersection information
content (25) is non-negative, we do not require an additional proof to show that the intersection entropy is non-negative. Moreover, the intersection entropy cannot be written as a Kullback-Leibler divergence.

Finally, similar to (26), we can decompose Eve's expected information into the following components,

$$
\begin{equation*}
H(X \sqcup Y)=H(X \sqcap Y)+H(X \backslash Y)+H(Y \backslash X) \tag{32}
\end{equation*}
$$

It is important to reiterate that, in contrast to (26), there is nothing which requires either of the two unique entropies to be zero. Thus, as shown in Figure 3, the Venn diagram which represents the union and intersection entropy differs from that which represents the union information content.

## 4. Synergistic Information Content

As discussed at the beginning of the previous section, and as required in Theorem 1, one of the defining features of Eve's information is that it is never greater than Johnny's information,

$$
\begin{equation*}
h(x, y) \geq h(x \sqcup y) \tag{33}
\end{equation*}
$$

Thus, we can compare how much more information Johnny has relative to Eve,

$$
\begin{equation*}
h(x \oplus y)=h(x, y)-h(x \sqcup y)=h(x, y)-\max (h(x), h(y))=\min (h(y \mid x), h(x \mid y)) \geq 0 \tag{34}
\end{equation*}
$$

This non-negative function is called the synergistic information content, and it quantifies how much more information one gets from knowing the joint probability $p_{X Y}(x, y)$ relative to merely knowing the marginal probabilities $p_{X}(x)$ and $p_{Y}(y)$. Figure 4 shows how this relationship can represented using a Venn diagram. Of course, by this definition, Johnny's information is equal to the union information content plus the synergistic information content, and hence, by using (26), we can decompose Johnny's information into the intersection information content, the unique information contents and the synergistic information contents,

$$
\begin{equation*}
h(x, y)=h(x \sqcup y)+h(x \oplus y)=h(x \sqcap y)+h(x \backslash y)+h(y \backslash x)+h(x \oplus y) \tag{35}
\end{equation*}
$$



Figure 4. (Left) This Venn diagram shows how the synergistic information $h(x \oplus y)$ can be defined by comparing the joint information content $h(x, y)$ from Figure 2 to the union information content $h(x \sqcup y)$ from Figure 3. Note that, for this particular realisation, we are assuming that $h(x)>h(y)$. It also provides a visual representation of the decomposition (40) of the joint information content $h(x, y)$. (Right) By rearranging the marginal entropies such that they match Figure 2 (albeit with different sizes here), it is easy to see why the mutual information content $i(x ; y)$ is equal to the intersection information content $h(x \sqcap y)$ minus the synergistic information content $h(x \oplus y)$, c.f. (38).

This decomposition can be seen in Figure 4, although it is important to recall that at least one of $h(x \backslash y)$ and $h(y \backslash x)$ must be equal to zero. In a similar manner, the extra information that Johnny has relative to Bob (13) can be decomposed into the unique information content from Alice and the
synergistic information content, and vice versa for the extra information that Johnny has relative to Alice (14),

$$
\begin{align*}
& h(x \mid y)=h(x \backslash y)+h(x \oplus y)  \tag{36}\\
& h(y \mid x)=h(y \backslash x)+h(x \oplus y) . \tag{37}
\end{align*}
$$

Now, recall that the mutual information content (18) is given by Indiana's information minus Johnny's information. By replacing Johnny's information with the union information content plus the synergistic information content via (34) and rearranging using (25), we get that the mutual information content is equal to the intersection information content minus the synergistic information content,

$$
\begin{equation*}
i(x ; y)=h(x)+h(y)-h(x, y)=h(x)+h(y)-h(x \sqcup y)-h(x \oplus y)=h(x \sqcap y)-h(x \oplus y) \tag{38}
\end{equation*}
$$

Indeed, this relationship can be identified in Figure 4. Clearly, the mutual information content is negative whenever the synergistic information content is greater than the intersection information content. From this perspective, the mutual information content can be negative because there is nothing to suggest that the synergistic information content should be no greater than the intersection information content. In other words, the additional surprise associated with knowing $p_{X Y}(x, y)$ relative to merely knowing $p_{X}(x)$ and $p_{Y}(y)$ can exceed the surprise of the least surprising marginal realisation.

Let us now quantify how much more information Johnny expects to get relative to Eve,

$$
\begin{equation*}
H(X \oplus Y)=\mathrm{E}_{X Y}[h(x \oplus y)]=H(X, Y)-H(X \sqcup Y) \geq 0 \tag{39}
\end{equation*}
$$

which we call the synergistic entropy. Crucially, although the synergistic information content is given by the minimum of the two conditional information contents, the synergistic entropy does not in general equal one of the two the conditional entropies. This is because, although Alice might observe the most surprising event in one joint realisation such that the synergistic information content is equal to Bob's information given Alice's information, Bob might observe the most surprising event in another realisation such that the synergistic information content is equal to Alice's information given Bob's information for that particular realisation. Thus, the synergistic entropy does not equal the conditional entropy for the same reason that unique entropies (29) and (30) can be simultaneously non-zero.

With the definition of synergistic entropy, it is not difficult to show that, similar to (35), the joint entropy can be decomposed into the following components,

$$
\begin{equation*}
H(X, Y)=H(X \sqcup \Upsilon)+H(X \oplus Y)=H(X \sqcap Y)+H(X \backslash Y)+H(Y \backslash X)+H(X \oplus Y) \tag{40}
\end{equation*}
$$

Figure 5 depicts this decomposition using a Venn diagram, and shows how the union entropy from Figure 3 is related to the joint entropy $H(X, Y)$. Likewise, similar to (36) and (37), it is easy to see that conditional entropies can be decomposed as follows,

$$
\begin{align*}
& H(X \mid Y)=H(X \backslash Y)+H(X \oplus Y)  \tag{41}\\
& H(Y \mid X)=H(Y \backslash X)+H(X \oplus Y) \tag{42}
\end{align*}
$$

Finally, as with (38), we can also show that the mutual information is equal to the intersection entropy minus the synergistic entropy,

$$
\begin{equation*}
I(X ; Y)=H(X \sqcap Y)-H(X \oplus Y) \geq 0 \tag{43}
\end{equation*}
$$

Although there is nothing to suggest that the synergistic information content must be no greater than the intersection information content, we know that the synergistic entropy must be no greater than the intersection entropy because $I(X ; Y) \geq 0$. In other words, the expected difference between the surprise
of the joint realisation and the most surprising marginal realisation cannot exceed the expected surprise of the least surprising realisation.


Figure 5. This Venn diagram shows how the synergistic entropy $H(X \oplus Y)$ can be defined by comparing the joint entropy $H(X, Y)$ from Figure 1 to the union entropy $H(X \sqcup Y)$ from Figure 3. It also provides a visual representation of the decomposition (40) of the joint entropy $H(X, Y)$.

## 5. Properties of the Union and Intersection Information Content

Theorem 1 determined the function form of Eve's information when Alice and Bob share their knowledge with her. We now wish to generalise this result to consider the situation whereby an arbitrary number of marginal observers share their information with Eve. Rather than try to directly determine the functional form, however, we proceed by considering the algebraic structure of shared marginal information.

If Alice and Bob observe the same realisation $x$ such that they have the same information $h(x)$, then upon sharing we would intuitively expect Eve to have the same information $h(x)$. Similarly, the minimum information that Eve could have received from either Alice or Bob should be the same information $h(x)$. Since the maximum and minimum operators are idempotent, the union and intersection information content both align with this intuition.

Property 1 (Idempotence). The union and intersection information content are idempotent,

$$
\begin{align*}
& h(x \sqcup x)=h(x),  \tag{44}\\
& h(x \sqcap x)=h(x) . \tag{45}
\end{align*}
$$

It also seems reasonable to expect that Eve's information should not depend on the order in which Alice and Bob share their information, nor should the minimum information that Eve could have received from either individual. Again, since the maximum and minimum operators are commutative, the union and intersection information content both align with our intuition.

Property 2 (Commutativity). The union and intersection information content are commutative,

$$
\begin{align*}
& h(x \sqcup y)=h(y \sqcup x),  \tag{46}\\
& h(x \sqcap y)=h(y \sqcap x) . \tag{47}
\end{align*}
$$

Now, suppose that Charlie is another individual who, similar to Alice and Bob, is separately observing some process, and let the random variable $Z$ represent her observations. Say that Dan is yet another individual with whom, similar to Eve, our observers can share their information. Intuitively, it should not matter whether Alice, Bob and Charlie share their information directly with Eve, or whether they share their information through Dan. To be specific, Alice and Bob could share their information with Dan such that his information is given by $h(x \sqcup y)$, and then Charlie
and Dan could subsequently share their information with Eve such that her information is given by $h((x \sqcup y) \sqcup z)$. Similarly, Bob and Charlie could share their information with Dan such that his information is given by $h(y \sqcup z)$, and then Alice and Dan could subsequently share their information with Eve such that her information is given by $h(x \sqcup(y \sqcup z))$. Alternatively, Alice, Bob and Charlie could entirely bypass Dan and share their information directly with Eve such that her information is given by $h(x \sqcup y \sqcup z)$. Since the maximum operator is associative, the union information content is the same in all three cases and hence aligns with our intuition. A similar argument can be made to show that the intersection information content is also associative.

Property 3 (Associative). The union and intersection information content are associative,

$$
\begin{align*}
& h(x \sqcup y \sqcup z)=h((x \sqcup y) \sqcup z)=h(x \sqcup(y \sqcup z))  \tag{48}\\
& h(x \sqcap y \sqcap z)=h((x \sqcap y) \sqcap z)=h(x \sqcap(y \sqcap z)) \tag{49}
\end{align*}
$$

Suppose now that Alice and Bob share their information with Dan such the information that he could have gotten from either Alice or Bob is given by $h(x \sqcap y)$. If Alice and Dan both share their information with Eve, then Eve's information is given by

$$
\begin{equation*}
h(x \sqcup(x \sqcap y))=\max (h(x), \min (h(x), h(y)))=h(x), \tag{50}
\end{equation*}
$$

and hence Bob's information has been absorbed by Alice's information. Now, suppose that Alice and Bob share their information with Dan such his information is given by $h(x \sqcup y)$. If Alice and Dan both share their information with Eve, then the information that Eve could have gotten from either Alice or Dan is given by

$$
\begin{equation*}
h(x \sqcap(x \sqcup y))=\min (h(x), \max (h(x), h(y)))=h(x) . \tag{51}
\end{equation*}
$$

Again, Bob's information has been absorbed by Alice's information. Both of these results are a consequence of the fact that the maximum and minimum operators are connected to each other by the absorption identity.

Property 4 (Absorption). The union and intersection information content are connected by absorption,

$$
\begin{align*}
& h(x \sqcup(x \sqcap y))=h(x),  \tag{52}\\
& h(x \sqcap(x \sqcup y))=h(x) . \tag{53}
\end{align*}
$$

Now, say that Daniella is, similar to Eve or Dan, an individual with whom our observers can share their information. Consider the following two cases: Firstly, suppose that Bob and Charlie share their information with Dan such that the information that Dan could have gotten from either Bob or Charlie is given by $h(y \sqcap z)$. If both Alice and Dan share their information with Eve, then her information is given by $h(x \sqcup(y \sqcap z))$. In the second case, suppose that Alice and Bob share their information with Dan such that his information is given by $h(x \sqcup y)$, while Alice and Charlie simultaneously share their information with Daniella such that her information is given by $h(x \sqcup z)$. If Dan and Daniella both share their information with Eve, then the information that she could have gotten from either Dan or Daniella is then given by $h((x \sqcup y) \sqcap(x \sqcup z))$. In both cases, Eve has the same information since the maximum operator is distributive,

$$
\begin{align*}
h(x \sqcup(y \sqcap z)) & =\max (h(x), \min (h(y), h(z))) \\
& =\min (\max (h(x), h(y)), \max (h(x), h(z)))=h((x \sqcup y) \sqcap(x \sqcup z)) . \tag{54}
\end{align*}
$$

Since the maximum and minimum operators are distributive over each other, regardless of whether Eve gets Alice's information and Bob's or Charlie's information, or if Eve gets Alice's and Bob's information or Alice's and Charlie's information, Eve has the same information. The same reasoning can be applied to show that, regardless of whether Eve gets Alice's information or Bob's and Charlie's
information, or if Eve gets Alice's or Bob's information and Alice's or Charlie's information, Eve has the same information.

Property 5 (Distributivity). The union and intersection information content are distribute over each other,

$$
\begin{align*}
& h(x \sqcup(y \sqcap z))=h((x \sqcup y) \sqcap(x \sqcup z)),  \tag{55}\\
& h(x \sqcap(y \sqcup z))=h((x \sqcap y) \sqcup(x \sqcap z)) . \tag{56}
\end{align*}
$$

Now, consider a set of $n$ individuals and let $X=\left\{X_{1}, X_{2}, \ldots, X_{n}\right\}$ be the joint random variable that represents their observations. Suppose that these individuals together observe the joint realisation $\boldsymbol{x}=\left\{x_{1}, x_{2}, \ldots, x_{n}\right\}$ from $\boldsymbol{X}$. By Property 3 and the general associativity theorem, it is clear that Eve's information is given by

$$
\begin{equation*}
h\left(x_{1} \sqcup x_{2} \sqcup \ldots \sqcup x_{n}\right)=\max \left(h\left(x_{1}\right), h\left(x_{2}\right), \ldots, h\left(x_{n}\right)\right) \geq 0, \tag{57}
\end{equation*}
$$

while the minimum information that Eve could have gotten from any individual observer is given by

$$
\begin{equation*}
h\left(x_{1} \sqcap x_{2} \sqcap \ldots \sqcap x_{n}\right)=\min \left(h\left(x_{1}\right), h\left(x_{2}\right), \ldots, h\left(x_{n}\right)\right) \geq 0 . \tag{58}
\end{equation*}
$$

This accounts for the situation whereby $n$ marginal observers directly share their information with Eve, and could clearly be considered for any subset $S$ of the observers $\boldsymbol{X}$. We now wish to consider all of the distinct ways that these marginal observers can share their information indirectly with Eve. As the following theorem shows, Properties 1-5 completely characterise the unique methods of marginal information sharing.

Theorem 2. The marginal information contents form a join semi-lattices $\langle\boldsymbol{x}, h(\sqcup)\rangle$ under the max operator. Separately, the marginal information contents form a meet semi-lattice $\langle\boldsymbol{x}, h(\sqcap)\rangle$ under the min operator.

Proof. Properties 1-3 completely characterise semi-lattices [16,17].
Theorem 3. The marginal information contents form a distributive lattice $\langle\boldsymbol{x}, h(\sqcup), h(\sqcap)\rangle$ under the max and min operators.

Proof. From Property 4, we have that the semi-lattices $\langle\boldsymbol{x}, h(\sqcup)\rangle$ and $\langle\boldsymbol{x}, h(\sqcap)\rangle$ are connected by absorption and hence form a lattice $\langle\boldsymbol{x}, h(\sqcup), h(\sqcap)\rangle$. By Property 5, this is a distributive lattice [16,17].

Each way that a set of $n$ observers can share their information with Eve such that she has distinct information corresponds to an element in partially ordered set, or more specifically the free distributive lattice on $n$ generators [16]. Figure 6 shows the free distributive lattices generated by $n=2$ and $n=3$ observers. The number of elements in this lattice is given by the ( $n$ )th Dedekind number (p. 273 [18]) (see also [19]). By the fundamental theorem of distributive lattices (or Birkhoff's representation theory), there is isomorphism between the union information content and set union, and between the intersection information content and set intersection [16,17,20,21]. It is this one-to-one correspondence that justifies our use of the terms union and intersection information content for $n$ variables in general. Every identity that holds in a lattice of sets will have a corresponding identity in this distributive lattice of information contents. Figure 6 also depicts the sets which correspond to each term in the lattice of information contents. Just as the cardinality of sets is non-decreasing as we consider moving up through the various terms in a lattice of sets, Eve's information is non-decreasing as we moving up through the various terms in the corresponding lattice of information contents. In particular, we can quantify the unique information content that Eve gets from one method of information sharing relative to any other method that is lower in the lattice.

Every property of the union and intersection information content that we have considered thus far has been directly inherited by the union and intersection entropy. However, there is one final property is not inherited by the entropies. If Alice and Bob share their information with Eve, then Eve's information is given by either Alice's or Bob's information, and similar for the information that Eve could have gotten from either Alice or Bob. As the subsequent theorem shows, this property enables us to greatly reduce the number of distinct terms in the distributive lattice for information content since any partially ordered set with a connex relation forms a total order.



$$
\begin{aligned}
a & =h((x \sqcup y) \sqcap(x \sqcup z) \sqcap(y \sqcup z)) \\
& =h((x \sqcup(y \sqcap z)) \sqcap(y \sqcup(x \sqcap z))) \\
& =h((x \sqcup(y \sqcap z)) \sqcap(z \sqcup(x \sqcap y))) \\
& =h((y \sqcup(x \sqcap z)) \sqcap(z \sqcup(x \sqcap y))) \\
& =h((y \sqcap(x \sqcup z)) \sqcup(z \sqcap(x \sqcup y))) \\
& =h((x \sqcap(y \sqcup z)) \sqcup(z \sqcap(x \sqcup y))) \\
& =h((x \sqcap(y \sqcup z)) \sqcup(y \sqcap(x \sqcup z))) \\
& =h((x \sqcap y) \sqcup(x \sqcap z) \sqcup(y \sqcap z)) \\
b & =h(y \sqcup(x \sqcap z))=h((x \sqcup y) \sqcap(y \sqcup z)) \\
c & =h(y \sqcap(x \sqcup z))=h((x \sqcap y) \sqcup(y \sqcap z))
\end{aligned}
$$




Figure 6. (Bottom right) The distributive lattices $\langle x, h(\sqcup), h(\sqcap)\rangle$ of information contents for two and three and three observers. It is also important to note that, by replacing $h, x, y$ and $z$ with $H, X, Y$ and $Z$, respectively, we can obtain the distributive lattices for entropy. In fact, this is crucial since Property 6 enables us to reduce the distributive lattice of information contents to a mere total order; however, this property does not apply to the entropies, and hence we cannot further simplify the lattice of entropies. (Top left) By the fundamental theorem of distributive lattices, the distributive lattices of marginal information contents has a one-to-one correspondence with the lattice of sets. Notice that the lattice for two sets corresponds to the Venn diagram for entropies in Figure 3.

Property 6 (Connexity). The union and intersection information content are given by at least one of

$$
\begin{equation*}
h(x \sqcup y)=h(x) \text { and } h(x \sqcap y)=h(y), \quad \text { or } \quad h(x \sqcup y)=h(y) \text { and } h(x \sqcap y)=h(x) \tag{59}
\end{equation*}
$$

## 6. Generalised Marginal Information Sharing

We now use Properties 1-6 to generalise the results of Theorem 1 and Section 3.
Theorem 4. The marginal information contents are a totally ordered set under the max and min operators.
Proof. A totally ordered set is a partially ordered set with the connex property (p. 2 [16]).
Figure 7 shows the totally ordered sets generated by $n=2$ and $n=3$ observers, and also depicts the corresponding sets. Although the number of distinct terms has been reduced, Eve's information is still non-decreasing as we move up through terms of the totally ordered set. If we now compare how much unique information Eve gets from a given method of information sharing relative to any other method of information sharing which is equal or lower in the totally ordered set, then we obtain a result which generalises (23) and (24) to consider more than two observers. Similarly, this total order enables us to generalise (25) using the maximum-minimum identity [22], which is a form of the principle of inclusion-exclusion [21] for a totally ordered set,

$$
\begin{align*}
h\left(x_{1} \sqcap x_{2} \sqcap \ldots \sqcap x_{n}\right) & =\min \left(h\left(x_{1}\right), h\left(x_{2}\right), \ldots, h\left(x_{n}\right)\right) \\
& =\sum_{k=1}^{n}(-1)^{k-1} \sum_{\substack{\boldsymbol{S} \subseteq X \\
|\boldsymbol{S}|=k}} \max \left(h\left(s_{1}\right), h\left(s_{2}\right), \ldots, h\left(s_{k}\right)\right) \\
& =\sum_{k=1}^{n}(-1)^{k-1} \sum_{\substack{\boldsymbol{S} \subseteq X \\
|\boldsymbol{S}|=k}} h\left(s_{1} \sqcup s_{2} \sqcup \ldots \sqcup s_{k}\right), \tag{60}
\end{align*}
$$

or, conversely,

$$
\begin{equation*}
h\left(x_{1} \sqcup x_{2} \sqcup \ldots \sqcup x_{n}\right)=\sum_{k=1}^{n}(-1)^{k-1} \sum_{\substack{\boldsymbol{S} \subseteq \boldsymbol{X} \\|\boldsymbol{S}|=k}} h\left(s_{1} \sqcap s_{2} \sqcap \ldots \sqcap s_{k}\right) . \tag{61}
\end{equation*}
$$



Figure 7. (Left) The total order of marginal information contents for two and three observers, whereby we have assumed that Alice's information $h(x)$ is greater than Bob's information $h(y)$, which is greater than Charlie's information $h(z)$. It is important to note that taking the expectation value over these information contents for each realisation, which may each have a different total orders, yields entropies which are merely partially ordered. It is for this reason that Property 6 does not apply to entropies. (Right) The Venn diagrams corresponding to the total order for for two and three observers and their corresponding information contents. Notice that the total order for two sets corresponds to the Venn diagram for information contents in Figure 3.

Now that we have generalised the union and intersection information content, similar to Section 3, let us now consider taking the expectation value for each term in the distributive lattice. For every joint realisation $x$ from $X$, there is a corresponding distributive lattice of information contents. Hence, similar to (27) and (29)-(31), we can consider taking the expectation value of each term in the lattice over all realisations. Since the expectation is a linear operator, this yields a set of entropies that are also idempotent, commutative, associative, absorptive and distributive, only now over the random variables from $\boldsymbol{X}$. Thus, the information that Eve expects to gain from a single realisation for a particular method of information sharing also corresponds to a term in a free distributive lattice generated by $n$. This distributive lattice for entropies can be seen in Figure 6 by replacing $x, y, z$ and $h$ with $X, Y, Z$ and $H$, respectively.

Crucially, however, Property 6 does not hold for the entropies-it is not true that Eve's expected information $H(X \sqcup Y)$ is given by either Alice's expected information $H(X)$ or Bob's expected information $H(Y)$. Thus, despite the fact that the distributive lattice of information content can be reduced to a total order, the distributive lattice of entropies remains partially ordered. Although the information contents are totally ordered for every realisation, this order is not in general the same for every realisation. Consequently, when taking the expectation value across many realisations to yield the corresponding entropies, the total order is not maintained, and hence we are left with a partially ordered set of entropies. Indeed, we already saw the consequences of this result in Figure 3 whereby Alice's and Bob's information content was totally ordered for any one realisation, but their expected information was partially ordered.

## 7. Multivariate Information Decomposition

In Section 4, we use the shared marginal information from Section 3 to decompose the joint information content into four distinct components. Our aim now is to use the generalised notion of shared information from the previous section to produce a generalised decomposition of the joint information content. To begin, suppose that Johnny observes the joint realisation $(x, y, z)$ while Alice, Bob and Charlie observe the marginal realisations $x, y$ and $z$, respectively, and say that Alice, Bob and Charlie share their information with Eve such that her information is given by $h(x \sqcup y \sqcup z)$. Clearly, Johnny has at least as much information as Eve,

$$
\begin{equation*}
h(x, y, z) \geq h(x \sqcup y \sqcup z) \tag{62}
\end{equation*}
$$

Thus, we can compare how much more information Johnny has relative to Eve,

$$
\begin{equation*}
h(x \oplus y \oplus z)=h(x, y, z)-h(x \sqcup y \sqcup z)=\min (h(y, z \mid x), h(x, z \mid y), h(x, y \mid z)) \geq 0 . \tag{63}
\end{equation*}
$$

This non-negative function generalises the earlier definition of the synergistic information content (34) such that it now quantifies how much information one gets from knowing the joint probability $p_{X Y Z}(x, y, z)$ relative to merely knowing the three marginal probabilities $p_{X}(x), p_{Y}(y)$ and $p_{Z}(z)$. Figure 8 shows how this relationship can be represented using a Venn diagram.

Now, consider three more observers, Joan, Jonas, and Joanna, who observe the joint marginal realisations $(x, y),(x, z)$ and $(y, z)$, respectively. Clearly, these additional observers greatly increase the number of distinct ways in which marginal information might be shared with Eve. For example, if Alice and Joanna share their information, then Eve's information is given by $h(x \sqcup(y, z))$. Alternatively, if Joan and Jonas share their information, then Eve's information is given by $h((x, y) \sqcup(x, z))$. Perhaps most interestingly, if Joan, Jonas and Joanna share their information, then Eve's information is given by $h((x, y) \sqcup(x, z) \sqcup(y, z))$. Moreover, we know that Johnny has at least as much information as Eve has in this situation,

$$
\begin{equation*}
h(x, y, z) \geq h((x, y) \sqcup(x, z) \sqcup(y, z)) \tag{64}
\end{equation*}
$$

Thus, by comparing how much more information Johnny has relative to Eve in this situation, we can define a new type of synergistic information content that quantifies how much information one gets from knowing the full joint realisation to merely knowing all of the pairwise marginal realisations,

$$
\begin{equation*}
h((x, y) \oplus(x, z) \oplus(y, z))=h(x, y, z)-h((x, y) \sqcup(x, z) \sqcup(y, z))=\min (h(z \mid x, y), h(y \mid x, z), h(x \mid y, z)) \tag{65}
\end{equation*}
$$

$$
\begin{aligned}
& h(z)=h(x \sqcap y \sqcap z)= \\
& h(x \sqcap z)=h(y \sqcap z)
\end{aligned}
$$

$$
h(x, y, z)
$$

$$
\begin{array}{ll}
h(y \backslash z)= \\
h(y \backslash(y \sqcap z))
\end{array} \begin{aligned}
& h(x \backslash y)= \\
& h(x \backslash(y \sqcup z))
\end{aligned}
$$

$$
\begin{array}{r}
i \\
--\stackrel{k}{v}
\end{array}
$$

$$
=h(x \sqcap y)
$$

$h(x \sqcup y)=h(x \sqcup z)$

Figure 8. Similar to Figure 4, this Venn diagram shows how the synergistic information $h(x \oplus y \oplus z)$ can be defined by comparing the joint information content $h(x, y, z)$ to the union information content $h(x \sqcup y \sqcup z)$. Note that, for this particular realisation, we are assuming that $h(x)>h(y)>h(z)$.

Of course, these new ways to share joint information are not just restricted to the union information. If Alice and Joanna share their information, then the information that Eve could have gotten from either is given by $h(x \sqcap(y, z))$. It is also worthwhile noting that this quantity is not less than the information that Eve could have gotten from either Alice's information or Bob's and Charlie's information,

$$
\begin{equation*}
h(x \sqcap(y, z)) \geq h(x \sqcap(y \sqcup z)) . \tag{66}
\end{equation*}
$$

Thus, we can also consider defining new types of synergistic information content associated with these this mixed type comparisons,

$$
\begin{equation*}
h(x \sqcap(y \oplus z))=h(x \sqcap(y, z))-h(x \sqcap(y \sqcup z)) . \tag{67}
\end{equation*}
$$

However, it is important to note that this quantity does not equal $\min (h(x), \min (h(z \mid y), h(y \mid z)))$.
With all of these new ways to share joint marginal information, it is not immediately clear how we should decompose Johnny's information. Nevertheless, let us begin by considering the algebraic structure of joint information content. From the inequality (12), we know that any pair of marginal information contents $h(x)$ and $h(y)$ are upper-bounded by the joint information content $h(x, y)$. It is also easy to see that the joint information content is idempotent, commutative and associative. Together, these properties are sufficient for establishing that the algebraic structure of joint information content is that of a join semi-lattice [16] which we denote by $\langle\boldsymbol{x} ; h()$,$\rangle . Figure 9$ shows the semi-lattices generated by $n=2$ and $n=3$ observers.

We now wish to establish the relationship between this semi-lattice of joint information content $\langle x ; h()$,$\rangle and the distributive lattice of shared marginal information \langle x ; h(\sqcup), h(\sqcap)\rangle$. In particular, since our aim is to decompose Johnny's information, consider the relationship between the join semi-lattice $\langle\boldsymbol{x} ; h()$,$\rangle and the meet semi-lattice \langle\boldsymbol{x} ; h(\sqcap)\rangle$, which is also depicted in Figure 9. In contrast to the semi-lattice of union information content $\langle x ; h(\sqcup)\rangle$, the semi-lattice $\langle x ; h()$,$\rangle is not connected to$ the semi-lattice $\langle x ; h(\sqcap)\rangle$. Although the intersection information content absorbs the joint information content, since

$$
\begin{equation*}
h(x \sqcap(x, y))=h(x) \tag{68}
\end{equation*}
$$

for all $h(x)$ and $h(y)$, the joint information content does not absorb the intersection information content since $h(x,(x \sqcap y))$ is equal to $h(x, y)$ for $h(x) \geq h(y)$, i.e., is not equal to $h(x)$ as required for absorption. Since the the join semi-lattice $\langle x ; h()$,$\rangle is not connected to the meet semi-lattice \langle x ; h(\sqcap)\rangle$ by absorption, their combined algebraic structure is not a lattice.


Figure 9. (Top-middle and left) The join semi-lattice $\langle\boldsymbol{h} ;()$,$\rangle for n=2$ and $n=3$ marginal observers. Johnny's information is always given by the joint information content at the top of the semi-lattice, while the information content of individuals such as Alice, Bob and Charlie who observe single realisations are found at the bottom of the semi-lattice. The information content of joint marginal observers such as Joanna, Jonas and Joan are found in between these two extremities. (Bottom-middle and right) The meet semi-lattice $\langle\boldsymbol{h} ; \Pi\rangle$ for $n=2$ and $n=3$ marginal observers. Since these two semi-lattices are not connect by absorption, their combined structure is not a lattice.

Despite the fact that the overall algebraic structure is not a lattice, there is a lattice sub-structure $\langle\mathcal{A}(\boldsymbol{x}), \preceq\rangle$ within the general structure. This substructure is isomorphic to the redundancy lattice from the partial information decomposition [23] (see also [24]), and its existence is a consequence of the fact that the intersection information content absorbs the joint information content in (68). To identify this lattice, we must first determine the reduced set of elements $\mathcal{A}(\boldsymbol{x})$ upon which it is defined. We begin by considering the set of all possible joint realisations which is given by $\mathcal{P}_{1}(\boldsymbol{x})$ where $\mathcal{P}_{1}(x)=\mathcal{P}(x) \backslash \varnothing$. Elements of this set $\mathcal{P}_{1}(x)$ correspond to the elements from the join semi-lattice $\langle\boldsymbol{x} ; h()$,$\rangle , e.g., the elements \{x\}$ and $\{x, y\}$ correspond to $h(x)$ and $h(x, y)$, respectively. In alignment with Williams and Beer [23], we call the elements of $\mathcal{P}_{1}(\boldsymbol{x})$ sources and denote them by $\boldsymbol{A}_{1}, \boldsymbol{A}_{2}, \ldots, \boldsymbol{A}_{k}$. Next, we consider set of all possible collections of sources which are given by the set $\mathcal{P}_{1}\left(\mathcal{P}_{1}(\boldsymbol{x})\right)$. Each collection of sources corresponds to an element of the meet semi-lattice $\left\langle\mathcal{P}_{1}(x) ; h(\square)\right\rangle$, or a particular way in which we can evaluate the intersection information content of a group of joint information contents. For example, the collections of sources $\{\{x\},\{y\}\}$ and $\{\{x\},\{y, z\}\}$ correspond to the $h(x \sqcap y)$ and $h(x \sqcap(y, z)))$, respectively. Not all of these collections of sources are distinct, however. Since the intersection information content absorbs the joint information content, we can remove the element $\{\{x\},\{x, y\}\}$ corresponding to $h(x \sqcap(x, y))$ as this information is already captured by the element $\{\{x\}\}$ corresponding to $h(x)$. In general, we can remove any collection of sources that corresponds to the intersection information content between a source $\boldsymbol{A}_{i}$ and any source $\boldsymbol{A}_{j}$ that is in the down-set $\downarrow \boldsymbol{A}_{i}$ with respect to the join semi-lattice $\langle\boldsymbol{x} ; h()$,$\rangle . (A definition of the down-set can be$ found in [17]. Informally, the down-set $\downarrow A$ is the set of all elements that precede $A$.) By removing all such collections of sources, we get the following reduced set of collections of sources,

$$
\begin{equation*}
\mathcal{A}(x)=\left\{\alpha \in \mathcal{P}_{1}\left(\mathcal{P}_{1}(x)\right): \forall A_{i}, A_{j} \in \alpha, A_{i} \not \subset A_{j}\right\} . \tag{69}
\end{equation*}
$$

Formally, this set corresponds to the set of antichains on the lattice $\left\langle\mathcal{P}_{1}(\boldsymbol{x}), \subseteq\right\rangle$, excluding the empty set [23].

Now that we have determined the elements upon which the lattice sub-structure is defined, we must show that they indeed form a lattice. Recall that when constructing the set $\mathcal{A}(\boldsymbol{x})$, we first
considered the ordered elements of the semi-lattice $\langle x ; h()$,$\rangle and then subsequently consider the$ ordered elements of the semi-lattice $\left\langle\mathcal{P}_{1}(x) ; h(\sqcap)\right\rangle$. Thus, we need to show that these two orders can be combined together into one new ordering relation over the set $\mathcal{A}(x)$. This can be done by extending the approach underlying the construction of the set $\mathcal{A}(x)$ to consider any pair of collections of sets $\alpha$ and $\beta$ from $\mathcal{A}(\boldsymbol{x})$. In particular, the collection of sets $\beta$ precedes the collection of sets $\alpha$ if and only if for every source $\boldsymbol{B}$ from $\beta$, there exists a source $\boldsymbol{A}$ from $\alpha$ such that $\boldsymbol{A}$ is in the down-set $\downarrow \boldsymbol{B}$ with respect to the join-semi-lattice $\langle x ; h()$,$\rangle , or formally,$

$$
\begin{equation*}
\forall \alpha, \beta \in \mathcal{A}(\boldsymbol{x}),(\alpha \preceq \beta \Longleftrightarrow \forall \boldsymbol{B} \in \beta, \exists \boldsymbol{A} \in \alpha, \boldsymbol{A} \subseteq \boldsymbol{B}) . \tag{70}
\end{equation*}
$$

The fact that $\langle\mathcal{A}(\boldsymbol{x}), \preceq\rangle$ forms a lattice was proved by Crampton and Loizou [25,26] where the corresponding lattice is denoted $\left\langle\mathcal{A}(X), \preceq^{\prime}\right\rangle$ in their notation. Furthermore, they showed that this lattice is isomorphic to the distributive lattices, and hence the number of elements in the set $\mathcal{A}(\boldsymbol{x})$ for $n$ marginal observers is also given by the ( $n$ )th Dedekind number (p. 273 [18]) (see also [19]). Crampton and Loizou [26] also provided the meet $\wedge$ and join $\vee$ operations for this lattice, which are given by

$$
\begin{align*}
& \alpha \wedge \beta=\alpha \sqcup \beta,  \tag{71}\\
& \alpha \vee \beta=\uparrow \alpha \cap \uparrow \beta, \tag{72}
\end{align*}
$$

where $\underline{\alpha}$ denotes the set of minimal elements of $\alpha$ with respect to the semi-lattice $\langle\boldsymbol{x} ; h()$,$\rangle . (A definition$ of the set of minimal elements can be found in [17]. Informally, $\underline{\alpha}$ is the set of sources of $\alpha$ that are not preceded by any other sources from $\alpha$ with respect to the semi-lattice $\langle\boldsymbol{x} ; h()$,$\rangle .) This lattice \langle\mathcal{A}(\boldsymbol{x}), \preceq\rangle$ is the aforementioned sub-structure that is isomorphic to the redundancy lattice from Williams and Beer [23]. However, as it is a lattice over information contents, it is actually equivalent to the specificity lattice from [27]. Figure 10 depicts the redundancy lattice of information contents for $n=2$ and $n=3$ marginal observers.

Similar to how Eve's information is non-decreasing as we move up through the terms of the distributive lattice of shared information, the redundancy lattice of information contents enables to see that, for example, the information that Eve could have gotten from either Alice or Joanna $h(x \sqcap(y, z))$ is no less than the information that Eve could have gotten from Alice or Bob $h(x \sqcap y)$. Thus, by taking the information $h(\alpha)$ associated with the collection of sources $\alpha$ from $\mathcal{A}(\boldsymbol{x})$ and subtracting from it the information $h\left(\alpha_{i}\right)$ associated with any collection of sources $\alpha_{i}$ from the down-set $\downarrow \alpha$, we can evaluate the unique information $h\left(\alpha \backslash \alpha_{i}\right)$ provided by $\alpha$ relative to $\alpha_{i}$. Moreover, as per Williams and Beer [23], we can derive a function that quantifies the partial information content $h_{\partial}(\alpha)$ associated with the collection of sources $\alpha$ that is not available in any of the collections of sources that are covered by $\alpha$. (The set of collections of sources that are covered by $\alpha$ is denoted $\alpha^{-}$. A definition of the covering relation is provided in [17]. Informally, $\alpha^{-}$is the set collections of sources that immediately precede $\alpha$.) Formally, this function corresponds to the Möbius inverse of $h$ on the redundancy lattice $\langle\mathcal{A}(\boldsymbol{x}), \preceq\rangle$, and can be defined implicitly by

$$
\begin{equation*}
h(\alpha)=\sum_{\beta \preceq \alpha} h_{\partial}(\beta) . \tag{73}
\end{equation*}
$$

By subtracting away the partial information terms that strictly precede $\alpha$ from both sides, it is easy to see that the partial information content $h_{\partial}(\alpha)$ can be calculated recursively from the bottom of the redundancy lattice of information contents,

$$
\begin{equation*}
h_{\partial}(\alpha)=h(\alpha)-\sum_{\beta \prec \alpha} h_{\partial}(\beta), \tag{74}
\end{equation*}
$$



Figure 10. (Top left) The redundancy lattices $\langle\mathcal{A}(x), \preceq\rangle$ of information contents for two and three and three observers. Each note in the lattice corresponds to an element in $\mathcal{A}(x)$ from (69), while the ordering between elements is given by $\preceq$ from (70). (Bottom right) The partial information contents $h_{\partial}(\alpha)$ corresponding to the redundancy lattices of information contents for two and three observers.

As the following theorem shows, the partial information content $h_{\partial}(\alpha)$ can be written in closed-form.
Theorem 5. The partial information content $h_{\partial}(\alpha)$ is given by

$$
\begin{align*}
h_{\partial}(\alpha) & =h(\alpha)-h\left(\alpha_{1}^{-} \sqcup \alpha_{2}^{-} \sqcup \ldots \sqcup \alpha_{\left|\alpha^{-}\right|}^{-}\right) \\
& =h(\alpha)-\max \left(h\left(\alpha_{1}^{-}\right), h\left(\alpha_{2}^{-}\right), \ldots, h\left(\alpha_{\left|\alpha^{-}\right|}^{-}\right)\right) \geq 0, \tag{75}
\end{align*}
$$

where each $\alpha_{i}^{-}$is a collection of sets from $\alpha^{-}$.
Proof. For $S \subseteq \mathcal{A}(x)$, define the set-additive function

$$
\begin{equation*}
f(\boldsymbol{S})=\sum_{\beta \in S} h_{\partial}(\beta) \tag{76}
\end{equation*}
$$

From (73), we have that $h(\alpha)=f(\downarrow \alpha)$. The partial information can then by subtracting the set additive on the down-set $\downarrow \alpha$ from the set additive function on the strict down-set $\downarrow \alpha$,

$$
\begin{equation*}
h_{\partial}(\alpha)=f(\downarrow \alpha)-f(\downarrow \alpha)=f(\downarrow \alpha)-f\left(\bigcup_{\beta \in \alpha^{-}} \downarrow \beta\right) \tag{77}
\end{equation*}
$$

By applying the principle of inclusion-exclusion [21], we get that

$$
\begin{equation*}
h_{\partial}(\alpha)=f(\downarrow \alpha)-\sum_{k=1}^{\left|\alpha^{-}\right|}(-1)^{k-1} \sum_{\substack{S \subseteq \alpha^{-} \\|\bar{S}|=k}} f\left(\bigcap_{\sigma \in S} \downarrow \sigma\right) \tag{78}
\end{equation*}
$$

For any lattice $L$ and $A \subseteq L$, we have that $\bigcap_{a \in A} \downarrow a$ is equal to $\downarrow(\Lambda A)$ (p. 57 [17]), and since the meet operation is given by the intersection information content, we have that

$$
\begin{align*}
h_{\partial}(\alpha) & =f(\downarrow \alpha)-\sum_{k=1}^{\left|\alpha^{-}\right|}(-1)^{k-1} \sum_{\substack{S \subseteq \alpha^{-} \\
|\bar{S}|=k}} h\left(s_{1} \sqcap s_{2} \sqcap \ldots \sqcap s_{k}\right) \\
& =h(\alpha)-h\left(\alpha_{1} \sqcup \alpha_{2} \sqcup \ldots \sqcup \alpha_{\left|\alpha^{-}\right|}\right), \tag{79}
\end{align*}
$$

where the final step has been made using (61) and (76).
The closed-form solution (75) from Theorem 5 is the same as the closed-form solution presented in Theorem A2 from Finn and Lizier [27]. This, together with the aforementioned fact that the lattice $\langle\mathcal{A}(x), \preceq\rangle$ is equivalent to the specificity lattice, means that each partial information content $h_{\partial}(\alpha)$ is equal to the partial specificity $i_{\partial}^{+}(\alpha \rightarrow t)$ from (A22) of [27]. As such, the partial information decomposition present in this paper is equivalent to the pointwise partial information decomposition presented in [27].

Let us now use the closed-form solution (75) from Theorem 5 to evaluate the partial information contents for the $n=2$ redundancy lattice of information contents. Starting from the bottom, we get the intersection information content,

$$
\begin{equation*}
h_{\partial}(x \sqcap y)=h(x \sqcap y) \tag{80}
\end{equation*}
$$

followed by the unique information contents,

$$
\begin{align*}
& h_{\partial}(x)=h(x)-h(x \sqcap y)=h(x \backslash y)  \tag{81}\\
& h_{\partial}(y)=h(y)-h(x \sqcap y)=h(y \backslash x) \tag{82}
\end{align*}
$$

and, finally, the synergistic information content,

$$
\begin{equation*}
h_{\partial}(x, y)=h(x, y)-h(x \sqcup y)=h(x \oplus y) \tag{83}
\end{equation*}
$$

It is clear that these partial information contents recover the intersection, unique and synergistic information contents from Sections 3 and 4. Moreover, by inserting these partial terms back into (73) for $\alpha=\{\{x, y\}\}$, we recover the earlier decomposition (35) of Johnny's information,

$$
\begin{equation*}
h(x, y)=h_{\partial}(x \sqcap y)+h_{\partial}(x)+h_{\partial}(y)+h_{\partial}(x, y)=h(x \sqcap y)+h(x \backslash y)+h(y \backslash x)+h(x \oplus y) \tag{84}
\end{equation*}
$$

Of course, our aim is to generalise this result such that we can decompose the joint information content for an arbitrary number of marginal realisations. This can be done by first evaluating the partial information contents over the redundancy lattice corresponding to $n$ marginal realisations, and then subsequently inserting the results back into (73) for $\alpha=\left\{\left\{x_{1}, x_{2}, \ldots, x_{n}\right\}\right\}$. For example, we can invert the $n=3$ redundancy lattice of information contents which yields the partial information contents shown in Figure 10. (The inversion is evaluated in the Appendix A.) When inserted back into (73), we get the following decomposition for Johnny's information,

$$
\begin{align*}
h(x, y, z)= & h(x \sqcap y \sqcap z) \\
& +h((x \sqcap y) \backslash z)+h((x \sqcap z) \backslash y)+h((y \sqcap z) \backslash x) \\
& +h(x \sqcap(y \oplus z))+h(y \sqcap(x \oplus z))+h(z \sqcap(x \oplus y)) \\
& +h(x \backslash(y, z))+h(y \backslash(x, z))+h(z \backslash(x, y))+h((x \oplus y) \sqcap(x \oplus z) \sqcap(y \oplus z))  \tag{85}\\
& +h((x \oplus y) \sqcap(x \oplus z) \backslash(y, z))+h((x \oplus y) \sqcap(y \oplus z) \backslash(x, z))+h((x \oplus z) \sqcap(y \oplus z) \backslash(x, y)) \\
& +h((x \oplus y) \backslash((x, z) \sqcup(y, z)))+h((x \oplus z) \backslash((x, y) \sqcup(y, z)))+h((y \oplus z) \backslash((x, y) \sqcup(x, z))) \\
& +h((x, y) \oplus(x, z) \oplus(y, z)) .
\end{align*}
$$

Finally, we can also consider taking the expectation value of each term in the redundancy lattice of information contents. Since the expectation is a linear and monotonic operator, the resulting expectation values will inherit the structure of the redundancy lattice of information contents and so form a redundancy lattice of entropies, i.e., Figure 10 with $x, y, z$ and $h$ replaced by $X, Y, Z$ and $H$, respectively. By inverting the $n=2$ redundancy lattice of entropies, we can recover the decomposition (40) from Figure 5. Furthermore, inverting the $n=3$ lattice generalises this result and is depicted in Figure 11.

$$
\begin{array}{ll}
b=H((X \oplus Y) \sqcap(X \oplus Z) \backslash(Y, Z)) & e=H((X \oplus Y) \backslash(X, Z) \sqcup(Y, Z)) \\
c=H((X \oplus Y) \sqcap(Y \oplus Z) \backslash(X, Z)) & f=H((X \oplus Z) \backslash(X, Y) \sqcup(Y, Z)) \\
d=H((X \oplus Z) \sqcap(Y \oplus Z) \backslash(X, Y)) & g=H((Y \oplus Z) \backslash(X, Y) \sqcup(X, Z))
\end{array}
$$



Figure 11. This Venn provides a visual representation of the decomposition of the joint entropy $H(X, Y, Z)$. This decomposition is given by replacing $x, y, z$ and $h$ with $X, Y, Z$ and $H$ in (85), respectively.

## 8. Union and Intersection Mutual Information

Suppose that Alice, Bob and Johnny are now additionally and commonly observing the variable $Z$. When a realisation $(x, y, z)$ occurs, Alice's information for $z$ is given by the conditional information content $h(x \mid z)$, while Bob's conditional information is given by $h(y \mid z)$ and Johnny's conditional information is given by $h(x, y \mid z)$. By using the same argument as in Section 3, it is easy to see that Eve's conditional information given $z$ is given by the conditional union information content,

$$
\begin{equation*}
h(x \sqcup y \mid z)=\max (h(x \mid z), h(y \mid z)) . \tag{86}
\end{equation*}
$$

Likewise, we can define the conditional unique information contents and conditional intersection information content, respectively,

$$
\begin{align*}
& h(x \backslash y \mid z)=h(x \sqcup y \mid z)-h(y \mid z)=\max (h(x \mid z)-h(y \mid z), 0),  \tag{87}\\
& h(y \backslash x \mid z)=h(x \sqcup y \mid z)-h(x \mid z)=\max (0, h(y \mid z)-h(x \mid z))  \tag{88}\\
& h(x \sqcap y \mid z)=h(x \mid z)+h(y \mid z)-h(x \sqcup y \mid z)=\min (h(x \mid z), h(y \mid z)) . \tag{89}
\end{align*}
$$

Furthermore, since Johnny's conditional information $h(x, y \mid z)$ is no less than Eve's conditional information content $h(x \sqcup y \mid z)$, we can also define the conditional synergistic information content,

$$
\begin{equation*}
h(x \oplus y \mid z)=h(x, y \mid z)-h(x \sqcup y \mid z)=\min (h(y \mid x, z), h(x \mid y, z)) . \tag{90}
\end{equation*}
$$

Similar to (35), we can decompose Johnny's conditional information $h(x, y \mid z)$ into the following components,

$$
\begin{equation*}
h(x, y \mid z)=h(x \sqcup y \mid z)+h(x \oplus y \mid z)=h(x \sqcap y \mid z)+h(x \backslash y \mid z)+h(y \backslash x \mid z)+h(x \oplus y \mid z) . \tag{91}
\end{equation*}
$$

Moreover, similar to (38), the conditional mutual information content is equal to the difference between the conditional intersection information content and the conditional synergistic information content,

$$
\begin{equation*}
i(x ; y \mid z)=h(x \mid z)+h(y \mid z)-h(x, y \mid z)=h(x \sqcap y \mid z)-h(x \oplus y \mid z) \tag{92}
\end{equation*}
$$

Notice that all of the above definitions directly correspond to the definitions of the unconditioned quantities, with all probability distributions conditioned on $z$ here.

Let us now consider how much information each of our observers have about the commonly observed realisation $z$. The information that Alice has about $z$ from observing $x$ is given by the mutual information content,

$$
\begin{equation*}
i(x ; z)=h(x)-h(x \mid z) \tag{93}
\end{equation*}
$$

Similarly, Bob's information about $z$ is given by $i(y ; z)$, while Johnny's information is given by the joint mutual information content $i(x, y ; z)$. Thus, the question naturally arises-are we able to quantify how much information Eve has about the realisation $z$ from knowing Alice's and Bob's shared information?

Clearly, we could consider defining the union mutual information content,

$$
\begin{equation*}
i(x \sqcup y ; z)=h(x \sqcup y)-h(x \sqcup y \mid z) . \tag{94}
\end{equation*}
$$

It is important to note that, while the mutual information can be defined in three different ways $i(x, z)=h(x)-h(x \mid z)=h(x)+h(z)-h(x, z)=h(z)-h(z \mid x)$, there is only one way in which one can define this function. (Indeed, this point aligns well with our argument based on exclusions presented in [28].) Similar to (94), we could consider respectively defining the unique mutual information contents, the intersection mutual information content and synergistic mutual information content,

$$
\begin{align*}
i(x \backslash y ; z) & =h(x \backslash y)-h(x \backslash y \mid z),  \tag{95}\\
i(y \backslash x ; z) & =h(y \backslash x)-h(y \backslash x \mid z),  \tag{96}\\
i(x \sqcap y ; z) & =h(x \sqcap y)-h(x \sqcap y \mid z),  \tag{97}\\
i(x \oplus y ; z) & =h(x \oplus y)-h(x \oplus y \mid z) \tag{98}
\end{align*}
$$

As with the mutual information content (18), there is nothing to suggest that these quantities are non-negative. Of course, the mutual information or expected mutual information content (20) is non-negative. Thus, with this in mind, consider defining the union mutual information

$$
\begin{equation*}
I(X \sqcup Y ; Z)=\mathrm{E}_{X Y Z}[i(x \sqcup y ; z)] \tag{99}
\end{equation*}
$$

However, there is nothing to suggest that this function is non-negative. Consequently, it is dubious to claim that this function represents Eve's expected information about $Z$, and is similarly fallacious to say that Eve's information about $z$ is given by the union mutual information content (94). Indeed, by inserting the definitions (21) and (86) into (94), it is easy to see why it is difficult to interpret these functions,

$$
\begin{align*}
i(x \sqcup y ; z) & =\max (h(x), h(y))-\max (h(x \mid z), h(y \mid z)) \\
& =\max (\min (i(x ; z), h(x)-h(y \mid z)), \min (h(y)-h(x \mid z), i(y ; z))) \tag{100}
\end{align*}
$$

That is, the union mutual information content can mix the information content provided by one realisation with the conditional information content provided by another. Thus, there is no guarantee
that this function's expected value will be non-negative. It is perhaps best to interpret this function as being a difference between two surprisals, rather than a function which represent information. Of course, similar to the multivariate mutual information (9), the union mutual information can be used a summary quantity provided one is careful not to misinterpret its meaning. The same is true for the unique mutual informations, intersection mutual information and synergistic mutual information, which we can similarly define,

$$
\begin{align*}
I(X \backslash Y ; Z) & =\mathrm{E}_{X Y Z}[i(x \backslash y ; z)]  \tag{101}\\
I(Y \backslash X ; Z) & =\mathrm{E}_{X Y Z}[i(y \backslash x ; z)]  \tag{102}\\
I(X \sqcap Y ; Z) & =\mathrm{E}_{X Y Z}[i(x \sqcap y ; z)]  \tag{103}\\
I(X \oplus Y ; Z) & =\mathrm{E}_{X Y Z}[i(x \oplus y ; z)] \tag{104}
\end{align*}
$$

Despite lacking the clear interpretation that we had for the information contents, these functions share a similar algebraic structure. For example, by using (35) and (91), we can decompose the mutual information content into the following components,

$$
\begin{equation*}
i(x, y ; z)=i(x \sqcap y ; z)+i(x \backslash y ; z)+i(y \backslash x ; z)+i(x \oplus y ; z) \tag{105}
\end{equation*}
$$

which is similar to the earlier decomposition of the joint entropy (35). Moreover, similar to (38), by using (38) and (92), we get that the multivariate mutual information content is given by the difference between the intersection mutual information content and the synergistic mutual information content,

$$
\begin{align*}
i(x ; y ; z) & =i(x ; y)-i(x ; y \mid z)=h(x \sqcap y)-h(x \oplus y)-h(x \sqcap y)+h(x \oplus y \mid z) \\
& =i(x \sqcap y ; z)-i(x \oplus y ; z) \tag{106}
\end{align*}
$$

Of course, since the expectation value is a linear operator, both of these results can be carried over to the joint mutual information. Hence, the mutual information can be decomposed into the following components,

$$
\begin{equation*}
I(X, Y ; Z)=I(X \sqcap Y ; Z)+I(X \backslash Y ; Z)+I(Y \backslash X ; Z)+I(X \oplus Y ; Z) \tag{107}
\end{equation*}
$$

while the the multivariate mutual information is equal to the intersection mutual information minus the synergistic mutual information,

$$
\begin{align*}
I(X ; Y ; Z) & =I(X ; Y)-I(X ; Y \mid Z)=H(X \sqcap Y)-H(X \oplus Y)-H(X \sqcap Y)+H(X \oplus Y \mid Z) \\
& =I(X \sqcap Y ; Z)-I(X \oplus Y ; Z) \tag{108}
\end{align*}
$$

This latter result aligns with Williams and Beer's prior result that the multivariate mutual information conflates redundant and synergistic information (Equation (14) [23]).

## 9. Conclusions

The main aim of this paper has been to understand and quantify the distinct ways that a set of marginal observers can share their information with some non-observing third party. To accomplish this objective, we examined the distinct ways in which two marginal observers, Alice and Bob, can share their information with the non-observing individual, Eve, and introduced several novel information-theoretic quantities: the union information content, which quantifies how much information Eve gets from the Alice and Bob; the intersection information content, which quantifies how much information Eve could have gotten from either Alice or Bob; and the unique information content, which quantifies how much information Eve gets from Alice relative to Bob, and vice versa. We then investigated the algebraic structure of these new measures of shared marginal information and showed that the structure of shared marginal information is that of a distributive
lattice. Next, by using the fundamental theorem of distributive lattices, we showed that these new measures are isomorphic to the various unions and intersections of sets. This isomorphism is similar to Yeung's correspondence between multivariate mutual information and signed measure [6,7]. However, in contrast to Yeung's correspondence, the measures of information content presented in this paper are non-negative and maintain a clear operational meaning regardless of the number of realisations or variables involved. (This is, of course, excepting the mutual information contents presented in Section 8, which are not non-negative.)

The appearance of a lattice structure within the context of information theory is by no means novel. Han [12] developed a lattice-theoretic description of the entropy over a Boolean lattice generated by a set of random variables. This lattice encapsulates all linear sums and differences of the basic information-theoretic quantities, i.e., entropy, conditional entropy, mutual information and conditional mutual information. Moreover, this lattice structure captures several of the existing multivariate generalisations of mutual information [29], including the aforementioned multivariate mutual information (9) (which is also known as the interaction information [10], amount of information [2] or co-information [11]), the total correlation [30] (which is also known as the multivariate constraint [31], multi-information [32] or integration [33]), the dual total correlation [12] (which is also known as binding information [34]) and the novel measure of multivariate mutual information defined by Chan et al. [29] (see Han [12] and Chan et al. [29] for further details). Similar to the lattice of shared marginal information content, Han's lattice is distributive-indeed, on a fundamental level, it is this algebraic structure that enables Yeung [6,7] to establish a correspondence with signed measure. Nevertheless, there two important differences to note between Han's information lattice and the lattice of shared marginal information content: Firstly, Han's lattice is based upon the entropies of random variables rather than the information content of realisations. In principle, there is no reason why one could not consider the information content of a Boolean lattice generated by a set of realisations (although the mutual information content would not be non-negative). Secondly, the Möbius inverse on Han's information lattice yields the multivariate mutual information (9), which is not non-negative. In contrast, the partial information contents (75) that result from the Möbius inversion of the lattice of shared marginal information content are non-negative. Thus, in contrast to the multivariate mutual information, the new measures of multivariate information presented in this paper maintain their operational meaning for any number of random variables.

Similar to Han, Shannon [35] introduced his own information lattice, although it is based upon the notion of common information. In comparison to Shannon's other work, this paper is not well recognised. Indeed, this common information was later independently proposed and studied by Gács and Körner [36]. Shannon's original paper is relatively brief; however, Li and Chong [37] expanded upon Shannon's discussion by formalising his argument in terms of $\sigma$-algebras and sample space partitions (see also [38]). To be specific, they described a random variable $X$ as "being-richer-than" another random variable $Y$ if the former's sample space partition is finer than the latter's sample space partition. Moreover, if their $\sigma$-algebras coincide, then two random variables are said to be informationally equivalent. This relation naturally forms a partial order over a set of random variables. For all $X$ and $Y$, the joint variable $(X, Y)$ is the poorest amongst all of the variables that are richer than both $X$ and $Y$. Conversely, one can define a random variable $Z$ that is the richest amongst all of the variables that are poorer than both $X$ and $Y$. The entropy of this common variable $Z$ defines the aforementioned common information. In contrast to the joint variable $(X, Y)$, it is relatively difficult to characterise the common variable $Z[36,37,39]$. Nevertheless, its existence is sufficient for the definition of Shannon's information lattice [35,37]. There are several features that distinguish this lattice from the lattice of shared marginal information. Firstly, similar to Han's information lattice, the joint entropy and common information are defined in terms of entire random variables, rather than the information content of realisations. Secondly, even if we were to restrict ourselves to the comparing Shannon's information lattice to the lattice of shared marginal entropy, the meet and join operations for these lattices are fundamentally different. We have already discussed the between their respective join
operations, i.e., the joint entropy and union entropy, in Sections 3 and 4 . If we consider their respective meet operations, we get the common information is relatively restrictive compared to the intersection entropy, due to the fact that the common information requires one to identify the common random variable $Z$. This follows from the fact that the intersection information is greater than or equal the mutual information (43), which is in turn greater than or equal to the common information [36]. Finally, in general, Shannon's information lattice is not distributive, nor is it even modular [35,37]. Thus, unlike the lattice of share marginal information or Han's information lattice, the fundamental theorem of distributive lattice is not applicable, and hence Shannon's information lattice does inherit any set-like identities.

The secondary objective of this paper has been to understand and demonstrate how we can use the measures of shared information content to decompose multivariate information. We began by comparing the union information content to the joint information content and used this comparison to define a measure of synergistic information content that captures how much more information a full joint observer, Johnny, has relative to an individual, Eve, who knows which joint realisation has occurred, but only knows the marginal distributions. We showed how one can use this measure, together with the measures of shared information content, to decompose the joint information content. We then compared the algebraic structure of joint information to the lattice structure of shared information, and showed how one can find the redundancy lattice from the partial information decomposition [23] embedded within this larger algebraic structure. More specifically, since this paper considers information contents, this redundancy lattice is actually same as the specificity lattice from pointwise partial information decomposition $[27,28]$. This observation connects the work presented in this paper to the existing body of theoretical literature on information decomposition [23,40-62], and its applications [63-86]. (For a brief summary of this literature, see [24].) Nevertheless, in contrast to the pointwise partial information decomposition [27,28], most of these approaches aim to decompose the average mutual information rather than the information content. The ability to decompose information content, and pointwise mutual information, provides a unique perspective on multivariate dependency.

To our knowledge, the only other approach that attempts to provide this pointwise perspective is due to Ince [87]. Ince's approach proposes a method of information decomposition based upon the entropy, but can be applied to the information content (or in Ince's terminology, the local entropy). Of particular relevance to this paper, Ince obtains a result that is equivalent to (38) whereby the mutual information content is equal to the redundant information content minus the synergistic information content (Equation (5) [87]). However, Ince's definition of redundant information content differs from that of the intersection information content in (38). To be specific, it is based upon the sign of the multivariate mutual information content (or pointwise co-information), which is interpreted as a measure of "the set-theoretic overlap" of multiple information contents (or local entropies) (p. 7 [87]). However, as discussed in Section 1, this set-theoretic interpretation of the multivariate mutual information (co-information) is problematic. To account for these difficulties, Ince disregards the negative values, defining the redundant information content to equal to the multivariate mutual information when it is positive, and to be zero otherwise.

There are several avenues of inquiry for which this research will yield new insights, particularly in complex systems, neuroscience and communications theory. For instance, these measures might be used to better understand and quantify distributed intrinsic computation [66,79]. It is well known that that dynamics of individual regions in the brain depend synergistically on multiple other regions; synergistic information content might provide a means to quantify such dependencies in neural data [69,77,88-90]. Furthermore, these measures might be helpful for quantifying the synergistic encodings used in network coding [7]. Finally, it is well-known that many biological traits are not dependent on any one gene, but rather are synergistically dependent on two or more genes, and the decomposed information provides a means to quantify the unique, redundant and synergistic dependencies between a trait and a set of genes [91-94].

Author Contributions: C.F. and J.T.L. conceived the idea; C.F. prepared the original draft; and J.T.L. reviewed and edited the draft. All authors have read and agreed to the published version of the manuscript.
Funding: JL was supported through the Australian Research Council DECRA grant DE160100630 and through The University of Sydney Research Accelerator (SOAR) Fellowship program.

Acknowledgments: The authors would like to thank Nathan Harding, Richard Spinney, Daniel Polani, Leonardo Novelli and Oliver Cliff for helpful discussions relating to this manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## Appendix A

We can use the closed-form (75) to evaluate the partial information contents for the $n=3$ redundancy lattice of information contents. Starting from the bottom and working up through the redundancy lattice, we have the following partial information contents:

$$
\begin{align*}
h_{\partial}(x \sqcap y \sqcap z)= & h(x \sqcap y \sqcap z) ;  \tag{A1}\\
h_{\partial}(x \sqcap y)= & h(x \sqcap y)-h(x \sqcap y \sqcap z) \\
= & h((x \sqcap y) \backslash z) ;  \tag{A2}\\
h_{\partial}(x \sqcap(y, z))= & h(x \sqcap(y, z))-h((x \sqcap y) \sqcup(x \sqcap z)) \\
= & h(x \sqcap(y, z))-h(x \sqcap(y \sqcup z)) \\
= & h(x \sqcap(y \oplus z)) ;  \tag{A3}\\
h_{\partial}(x)= & h(x)-h(x \sqcap(y, z)) \\
= & h(x \backslash(y, z)) ;  \tag{A4}\\
h_{\partial}((x, y) \sqcap(x, z) \sqcap(y, z))= & h((x, y) \sqcap(x, z) \sqcap(y, z))-h((x \sqcap(y, z)) \sqcup(y \sqcap(x, z)) \sqcup(z \sqcap(x, y))) \\
= & h((x, y) \sqcap(x, z) \sqcap(y, z))-h(x \sqcap(y \oplus z))-h(y \sqcap(x \oplus z))-h(z \sqcap(x \oplus y)) \\
& \quad-h((x \sqcap y) \backslash z)-h((x \sqcap z) \backslash y)-h((y \sqcap z) \backslash x)-h(x \sqcap y \sqcap z) \\
= & h((x, y) \sqcap(x, z) \sqcap(y, z))-h(x \sqcap(y, z))-h(y \sqcap(x \oplus z))-h(z \sqcap(x \oplus y)) \\
& \quad-h((y \sqcap z) \backslash x)  \tag{ByLemmaA9.}\\
= & h((x \oplus y) \sqcap(x \oplus z) \sqcap(y \oplus z)) ;  \tag{A5}\\
& \quad+h(x \sqcap((x, y) \sqcap(x, z) \sqcap(y, z))) \\
h_{\partial}((x, y) \sqcap(x, z))= & h((x, y) \sqcap(x, z))-h(x \sqcup((x, y) \sqcap(x, z) \sqcap(y, z))) \\
= & h((x, y) \sqcap(x, z))-h(x)-h((x, y) \sqcap(x, z) \sqcap(y, z)) \\
= & h(((x, y) \sqcap(x, z)) \backslash(y, z))-h(x \backslash(y, z)) \\
= & h((x \oplus y) \sqcap(x \oplus z) \backslash(y, z)) ;  \tag{ByLemmaA2.}\\
h_{\partial}(x, y)= & h(x, y)-h(((x, y) \sqcap(x, z)) \sqcup((x, y) \sqcap(y, z)))  \tag{A6}\\
= & h(x, y)-h((x, y) \sqcap((x, z) \sqcup(y, z))) \\
= & h((x, y) \backslash((x, z) \sqcup(y, z))) \\
= & h((x \oplus y) \backslash((x, z) \sqcup(y, z))) ;
\end{align*}
$$

$$
\begin{align*}
h_{\partial}(x, y, z) & =h(x, y, z)-h((x, y) \sqcup(x, z) \sqcup(y, z)) \\
& =h((x, y) \oplus(x, z) \oplus(y, z)) . \tag{A8}
\end{align*}
$$

Lemma A1. We have the following identity,

$$
\begin{equation*}
h((x \sqcap y) \backslash z)=h((x \backslash z) \sqcap(y \backslash z)) . \tag{A9}
\end{equation*}
$$

Proof. From (23) and (25), we have that

$$
\begin{align*}
h((x \sqcap y) \backslash z) & =\max (\min (h(x), h(y))-h(z), 0)=\max (\min (h(x)-h(z), h(y)-h(z)), 0) \\
& =\min (\max (h(x)-h(z), 0), \max (h(y)-h(z), 0))=h((x \backslash z) \sqcap(y \backslash z)), \tag{A10}
\end{align*}
$$

where we have used the fact that $\max (a, \min (b, c))$ is equal to $\min (\max (a, b), \max (a, c))$.
Lemma A2. We have the following identity,

$$
\begin{equation*}
h(((x, y) \backslash(y, z)) \sqcap((x, z) \backslash(y, z)))=h(x \backslash(y, z))+h((x \oplus y) \sqcap(x \oplus z) \backslash(y, z)) \tag{A11}
\end{equation*}
$$

Proof. From (35), we have that

$$
\begin{align*}
h((x, y) \backslash(y, z)) & =h((x \sqcap y) \backslash(y, z))+h((x \backslash y) \backslash(y, z))+h((y \backslash x) \backslash(y, z))+h((x \oplus y) \backslash(y, z)) \\
& =h(x \backslash(y, z))+h((y \backslash x) \backslash(y, z))+h((x \oplus y) \backslash(y, z)), \tag{A12}
\end{align*}
$$

and since

$$
\begin{equation*}
h((y \backslash x) \backslash(y, z))=\max (h(y \backslash x)-h(y, z), 0)=0, \tag{A13}
\end{equation*}
$$

we get that

$$
\begin{equation*}
h((x, y) \backslash(y, z))=h(x \backslash(y, z))+h((x \oplus y) \backslash(y, z)) . \tag{A14}
\end{equation*}
$$

Finally, by inserting this result into (25), we get that

$$
\begin{aligned}
h(((x, y) \backslash(y, z)) & \sqcap((x, z) \backslash(y, z)) \\
& =\min (h(x \backslash(y, z))+h((x \oplus y) \backslash(y, z)), h(x \backslash(y, z))+h((x \oplus z) \backslash(y, z))) \\
& =h(x \backslash(y, z))+h(((x \oplus y) \backslash(y, z)) \sqcap((x \oplus z) \backslash(y, z))) .
\end{aligned}
$$

Lemma A3. We have the following identity,

$$
\begin{equation*}
h(((x, y) \backslash((x, z) \sqcup(y, z)))=h(((x \oplus y) \backslash((x, z) \sqcup(y, z))) . \tag{A15}
\end{equation*}
$$

Proof. By (35), we have that

$$
\begin{align*}
h((x, y) \backslash((x . z) \sqcup(y, z)))= & h((x \sqcap y) \backslash((x, z) \sqcup(y, z)))+h((x \backslash y) \backslash((x, z) \sqcup(y, z))) \\
& +h((y \backslash x) \backslash((x, z) \sqcup(y, z)))+h((x \oplus y) \backslash((x, z) \sqcup(y, z))) . \tag{A16}
\end{align*}
$$

Since $h(x \sqcap y), h(x \backslash y)$ and $h(y \backslash x)$ are less than $h(((x, z) \sqcup(y, z))$, from (23) we have that $h((x \sqcap y) \backslash$ $((x, z) \sqcup(y, z))), h((x \backslash y) \backslash((x, z) \sqcup(y, z)))$ and $h((y \backslash x) \backslash((x, z) \sqcup(y, z)))$ are all equal to 0 .

Lemma A4. We have the following identity,

$$
\begin{equation*}
h(x \sqcap(y \backslash x))=0 . \tag{A17}
\end{equation*}
$$

Proof. We have that

$$
\begin{equation*}
h(x \sqcap(y \backslash x))=h(x \sqcap(x, y))-h(x \sqcap x)=h(x \sqcap x)-h(x)=0, \tag{A18}
\end{equation*}
$$

where we have used (45) and (68).
Lemma A5. We have the following identity,

$$
\begin{equation*}
h((y \backslash x) \sqcap(y, z))=h(y \backslash x) . \tag{A19}
\end{equation*}
$$

Proof. We have that

$$
\begin{equation*}
h((y \backslash x) \sqcap(y, z))=h((y \backslash x) \sqcap(y \backslash x, y \sqcap x, z)) . \tag{A20}
\end{equation*}
$$

Using (68), this then reduces to

$$
\begin{equation*}
h((y \backslash x) \sqcap(y, z))=h(y \backslash x) . \tag{A21}
\end{equation*}
$$

Lemma A6. We have the following identity,

$$
\begin{equation*}
h(x \sqcap(x \oplus z))=0 . \tag{A22}
\end{equation*}
$$

Proof. We have

$$
h(x \sqcap(x \oplus z))=h(x \sqcap(x, z))-h(x \sqcap x)+h(x \sqcap(z \backslash x)),
$$

which then reduces via (68), (45) and Lemma A4 to:

$$
\begin{align*}
h(x \sqcap(x \oplus z)) & =h(x)-h(x)+h(x \sqcap(z \backslash x))+0  \tag{A23}\\
& =0 . \quad \square
\end{align*}
$$

Lemma A7. We have the following identity,

$$
\begin{equation*}
h((y \backslash x) \sqcap(x \oplus z))=h(y \sqcap(x \oplus z)) . \tag{A24}
\end{equation*}
$$

Proof. We have that

$$
h(y \sqcap(x \oplus z))=h((y \backslash x) \sqcap(x \oplus z))+h((y \sqcap x) \sqcap(x \oplus z)) .
$$

However, since $h(x \sqcap(x \oplus z))=0$ via Lemma A6 and therefore $h((y \sqcap x) \sqcap(x \oplus z))=0$, we have

$$
h(y \sqcap(x \oplus z))=h((y \backslash x) \sqcap(x \oplus z))
$$

as required.
Lemma A8. We have the following identity,

$$
\begin{equation*}
h((x \oplus y) \sqcap(x, z) \sqcap(y, z))=h(z \sqcap(x \oplus y))+h((x \oplus y) \sqcap(x \oplus z) \sqcap(y \oplus z)) . \tag{A25}
\end{equation*}
$$

Proof. We have that

$$
\begin{align*}
h((x \oplus y) \sqcap(x, z) \sqcap(y, z))= & h((x \oplus y) \sqcap x \sqcap(y, z))+h((x \oplus y) \sqcap(z \backslash x) \sqcap(y, z)) \\
& +h((x \oplus y) \sqcap(x \oplus z) \sqcap(y, z)) . \tag{A26}
\end{align*}
$$

Then, by using Lemma A6, we get that

$$
\begin{align*}
h((x \oplus y) \sqcap(x, z) \sqcap(y, z))= & h((x \oplus y) \sqcap(z \backslash x) \sqcap(y, z))+h((x \oplus y) \sqcap(x \oplus z) \sqcap(y, z)) \\
= & h((x \oplus y) \sqcap(z \backslash x) \sqcap y)+h((x \oplus y) \sqcap(z \backslash x) \sqcap(z \backslash y)) \\
& \quad+h((x \oplus y) \sqcap(z \backslash x) \sqcap(y \oplus z))+h((x \oplus y) \sqcap(x \oplus z) \sqcap y)  \tag{A27}\\
& \quad+h((x \oplus y) \sqcap(x \oplus z) \sqcap(z \backslash y))+h((x \oplus y) \sqcap(x \oplus z) \sqcap(y \oplus z)) \\
= & h((x \oplus y) \sqcap(z \backslash x) \sqcap(z \backslash y)+h((x \oplus y) \sqcap(x \oplus z) \sqcap(y \oplus z)),
\end{align*}
$$

where we have used Lemma A6 four more times. Finally, using Lemma A7, we get that

$$
\begin{equation*}
h((x \oplus y) \sqcap(x, z) \sqcap(y, z))=h((x \oplus y) \sqcap z)+h((x \oplus y) \sqcap(x \oplus z) \sqcap(y \oplus z)) \tag{A28}
\end{equation*}
$$

as required.
Lemma A9. We have the following identity,

$$
\begin{align*}
& h((x, y) \sqcap(x, z) \sqcap(y, z))=h( x \sqcap(y, z))+h(y \sqcap(x \oplus z))+h(z \sqcap(x \oplus y))+h((y \sqcap z) \backslash x) \\
&+h((x \oplus y) \sqcap(x \oplus z) \sqcap(y \oplus z)) . \tag{A29}
\end{align*}
$$

Proof. We have that

$$
\begin{align*}
h((x, y) \sqcap(x, z) \sqcap(y, z)) & =h(x \sqcap(x, z) \sqcap(y, z))+h((y \backslash x) \sqcap(x, z) \sqcap(y, z))+h((x \oplus y) \sqcap(x, z) \sqcap(y, z)) \\
& =h(x \sqcap(y, z))+h((y \backslash x) \sqcap(x, z))+h((x \oplus y) \sqcap(x, z) \sqcap(y, z)), \tag{A30}
\end{align*}
$$

where we have used (68) and Lemma A5. Next, we have that

$$
\left.\begin{array}{rl}
h((x, y) \sqcap(x, z) \sqcap(y, z))= & h(x
\end{array}\right) \quad
$$

where we have used Lemma A1, Lemma A4 and Lemma A7. Finally, we have that

$$
\begin{align*}
h((x, y) \sqcap(x, z) \sqcap(y, z))=h(x & \sqcap(y, z))+h((y \sqcap z) \backslash x)+h(y \sqcap(x \oplus z))+h(z \sqcap(x \oplus y)) \\
& +h((x \oplus y) \sqcap(x \oplus z) \sqcap(y \oplus z)), \tag{A32}
\end{align*}
$$

where we have used via Lemma A8.

## References

1. Reza, F. An Introduction to Information Theory, International student edition; McGraw-Hill: New York, NY, USA, 1961.
2. Kuo Ting, H. On the Amount of Information. Theory Probab. Appl. 1962, 7, 439-447. doi:10.1137/1107041. (Originally in Russian; English translation by D. Lieberman) [CrossRef]
3. Abramson, N. Information Theory and Coding; McGraw-Hill: New York, NY, USA, 1963.
4. Campbell, L. Entropy as a measure. IEEE Trans. Inf. Theory 1965, 11, 112-114. doi:10.1109/TIT.1965.1053712. [CrossRef]
5. Csiszar, I.; Körner, J. Information Theory: Coding Theorems for Discrete Memoryless Systems; Academic Press, Inc.: Cambridge, MA, USA, 1981.
6. Yeung, R.W. A new outlook on Shannon's information measures. IEEE Trans. Inf. Theory 1991, 37, 466-474. doi:10.1109/18.79902. [CrossRef]
7. Yeung, R.W. Information Theory and Network Coding; Springer Science \& Business Media: Berlin/Heidelberg, Germany, 2008.
8. MacKay, D.J. Information Theory, Inference and Learning Algorithms; Cambridge University Press: Cambridge, UK, 2003.
9. Fano, R.M. Transmission of Information: A Statistical Theory of Communication; M.I.T. Press: Cambridge, MA, USA, 1961.
10. McGill, W. Multivariate information transmission. Trans. IRE Prof. Group Inf. Theory 1954, 4, 93-111. [CrossRef]
11. Bell, A.J. The co-information lattice. In Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA, Granada, Spain, 22-24 September 2004; Volume 2003.
12. Han, T.S. Linear dependence structure of the entropy space. Inf. Control 1975, 29, 337-368. [CrossRef]
13. Cover, T.M.; Thomas, J.A. Elements of Information Theory; John Wiley \& Sons: Hoboken, NJ, USA, 2012.
14. Fano, R.M. The statistical theory of information. Il Nuovo Cimento 1959, 13, 353-372. doi:10.1007/BF02724671. [CrossRef]
15. Pinsker, M.S. Information and Information Stability of Random Variables and Processes; Holden-Day: San Francisco, CA, USA, 1964.
16. Grätzer, G. General Lattice Theory; Springer Science \& Business Media: Berlin/Heidelberg, Germany, 2002.
17. Davey, B.A.; Priestley, H.A. Introduction to Lattices and Order; Cambridge University Press: Cambridge, UK, 2002.
18. Comtet, L. Advanced Combinatorics: The Art of Finite and Infinite Expansions; Springer Science \& Business Media: Berlin/Heidelberg, Germany, 2012.
19. Inc., O.F. The On-Line Encyclopedia of Integer Sequences. 2019. Available online: https:/ / oeis.org/A000372 (accessed on 14 February 2020).
20. Birkhoff, G. Lattice Theory; American Mathematical Soc.: Providence, RI, USA, 1940; Volume 25.
21. Stanley, R.P. Enumerative Combinatorics; Cambridge University Press: Cambridge, UK, 1997; Volume 1,
22. Sheldon, R. A first Course in Probability; Pearson Education India: Delhi, India, 2002.
23. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information. arXiv 2010, arXiv:1004.2515.
24. Lizier, J.T.; Bertschinger, N.; Jost, J.; Wibral, M. Information Decomposition of Target Effects from Multi-Source Interactions: Perspectives on Previous, Current and Future Work. Entropy 2018, $20,307$. [CrossRef]
25. Crampton, J.; Loizou, G. The completion of a poset in a lattice of antichains. Int. Math. J. 2001, 1, 223-238.
26. Crampton, J.; Loizou, G. Two Partial Orders on the Set of Antichains. 2000. Available online: http: / /learninglink.ac.uk/oldsite/research/techreps/2000/bbkcs-00-09.pdf (accessed on 14 February 2020).
27. Finn, C.; Lizier, J.T. Pointwise partial information decomposition using the specificity and ambiguity lattices. Entropy 2018, 20, 297. [CrossRef]
28. Finn, C.; Lizier, J.T. Probability Mass Exclusions and the Directed Components of Mutual Information. Entropy 2018, 20, 826. [CrossRef]
29. Chan, C.; Al-Bashabsheh, A.; Ebrahimi, J.B.; Kaced, T.; Liu, T. Multivariate mutual information inspired by secret-key agreement. Proc. IEEE 2015, 103, 1883-1913. [CrossRef]
30. Watanabe, S. Information theoretical analysis of multivariate correlation. IBM J. Res. Dev. 1960, 4, 66-82. [CrossRef]
31. Garner, W.R. Uncertainty and structure as psychological concepts. Science 1963, 140, 799.
32. Studenỳ, M.; Vejnarová, J. The multiinformation function as a tool for measuring stochastic dependence. In Learning in Graphical Models; Springer: Berlin/Heidelberg, Germany, 1998; pp. 261-297.
33. Tononi, G.; Sporns, O.; Edelman, G.M. A measure for brain complexity: Relating functional segregation and integration in the nervous system. Proc. Natl. Acad. Sci. USA 1994, 91, 5033-5037. [CrossRef] [PubMed]
34. Abdallah, S.A.; Plumbley, M.D. A measure of statistical complexity based on predictive information with application to finite spin systems. Phys. Lett. A 2012, 376, 275-281. [CrossRef]
35. Shannon, C. The lattice theory of information. Trans. IRE Prof. Group Inf. Theory 1953, 1, 105-107. [CrossRef]
36. Gács, P.; Körner, J. Common information is far less than mutual information. Probl. Control. Inf. Theory 1973, 2, 149-162.
37. Li, H.; Chong, E.K. On a connection between information and group lattices. Entropy 2011, 13, 683-708. [CrossRef]
38. Yu, H.; Mineyev, I.; Varshney, L.R. A group-theoretic approach to computational abstraction: Symmetry-driven hierarchical clustering. arXiv 2018, arXiv:1807.11167.
39. Wolf, S.; Wultschleger, J. Zero-error information and applications in cryptography. In Proceedings of the IEEE Information Theory Workshop, San Antonio, TX, USA, 24-29 October 2004; pp. 1-6.
40. 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; Springer: Berlin/Heidelberg, Germany, 2013; pp. 251-269.
41. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. Phys. Rev. E 2013, 87, 012130. [CrossRef] [PubMed]
42. Harder, M. Information Driven Self-Organization of Agents and Agent Collectives. Ph.D. Thesis, University of Hertfordshire, Hatfield, UK, 2013.
43. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. Entropy 2014, 16, 2161-2183. [CrossRef]
44. Griffith, V.; Koch, C. Quantifying Synergistic Mutual Information. In Guided Self-Organization: Inception; Emergence, Complexity and Computation; Prokopenko, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 9, pp. 159-190.
45. Griffith, V.; Chong, E.K.; James, R.G.; Ellison, C.J.; Crutchfield, J.P. Intersection information based on common randomness. Entropy 2014, 16, 1985-2000. [CrossRef]
46. 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; pp. 2232-2236.
47. Barrett, A.B. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. Phys. Rev. E 2015, 91, 052802. [CrossRef]
48. Griffith, V.; Ho, T. Quantifying redundant information in predicting a target random variable. Entropy 2015, 17, 4644-4653. [CrossRef]
49. Olbrich, E.; Bertschinger, N.; Rauh, J. Information decomposition and synergy. Entropy 2015, 17, 3501-3517. [CrossRef]
50. Perrone, P.; Ay, N. Hierarchical Quantification of Synergy in Channels. Front. Robot. AI 2016, 2, 35. [CrossRef]
51. Rosas, F.; Ntranos, V.; Ellison, C.J.; Pollin, S.; Verhelst, M. Understanding interdependency through complex information sharing. Entropy 2016, 18, 38. [CrossRef]
52. Faes, L.; Marinazzo, D.; Stramaglia, S. Multiscale information decomposition: exact computation for multivariate Gaussian processes. Entropy 2017, 19, 408. [CrossRef]
53. Ince, A.A.R. Measuring multivariate redundant information with pointwise common change in surprisal. Entropy 2017, 19, 318. [CrossRef]
54. James, R.G.; Crutchfield, J.P. Multivariate dependence beyond shannon information. Entropy 2017, 19, 531. [CrossRef]
55. Kay, J.W.; Ince, R.A.; Dering, B.; Phillips, W.A. Partial and Entropic Information Decompositions of a Neuronal Modulatory Interaction. Entropy 2017, 19, 560. [CrossRef]
56. Makkeh, A.; Theis, D.O.; Vicente, R. Bivariate Partial Information Decomposition: The Optimization Perspective. Entropy 2017, 19, 530. [CrossRef]
57. Pica, G.; Piasini, E.; Chicharro, D.; Panzeri, S. Invariant components of synergy, redundancy, and unique information among three variables. Entropy 2017, 19, 451. [CrossRef]
58. Quax, R.; Har-Shemesh, O.; Sloot, P. Quantifying synergistic information using intermediate stochastic variables. Entropy 2017, 19, 85. [CrossRef]
59. Rauh, J.; Banerjee, P.; Olbrich, E.; Jost, J.; Bertschinger, N. On extractable shared information. Entropy 2017, 19, 328. [CrossRef]
60. Rauh, J.; Banerjee, P.; Olbrich, E.; Jost, J.; Bertschinger, N.; Wolpert, D. Coarse-graining and the Blackwell order. Entropy 2017, 19, 527. [CrossRef]
61. Rauh, J. Secret sharing and shared information. Entropy 2017, 19, 601. [CrossRef]
62. James, R.G.; Emenheiser, J.; Crutchfield, J.P. Unique information via dependency constraints. J. Phys. Math. Theor. 2018, 52, 014002. [CrossRef]
63. Williams, P.L.; Beer, R.D. Generalized measures of information transfer. arXiv 2011, arXiv:1102.1507.
64. Flecker, B.; Alford, W.; Beggs, J.M.; Williams, P.L.; Beer, R.D. Partial information decomposition as a spatiotemporal filter. Chaos Interdiscip. J. Nonlinear Sci. 2011, 21, 037104. [CrossRef] [PubMed]
65. Stramaglia, S.; Wu, G.R.; Pellicoro, M.; Marinazzo, D. Expanding the transfer entropy to identify information circuits in complex systems. Phys. Rev. E 2012, 86, 066211. [CrossRef] [PubMed]
66. 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, 15-19 April 2013; pp. 43-51.
67. Stramaglia, S.; Cortes, J.M.; Marinazzo, D. Synergy and redundancy in the Granger causal analysis of dynamical networks. New J. Phys. 2014, 16, 105003. [CrossRef]
68. Timme, N.; Alford, W.; Flecker, B.; Beggs, J.M. Synergy, redundancy, and multivariate information measures: An experimentalist's perspective. J. Comput. Neurosci. 2014, 36, 119-140. [CrossRef] [PubMed]
69. Wibral, M.; Lizier, J.T.; Priesemann, V. Bits from brains for biologically inspired computing. Front. Robot. 2015, 2, 5. [CrossRef]
70. Biswas, A.; Banik, S.K. Redundancy in information transmission in a two-step cascade. Phys. Rev. E 2016, 93, 052422. [CrossRef]
71. Frey, S.; Williams, P.L.; Albino, D.K. Information encryption in the expert management of strategic uncertainty. arXiv 2016, arXiv:1605.04233.
72. Timme, N.M.; Ito, S.; Myroshnychenko, M.; Nigam, S.; Shimono, M.; Yeh, F.C.; Hottowy, P.; Litke, A.M.; Beggs, J.M. High-degree neurons feed cortical computations. PLoS Comput. Biol. 2016. [CrossRef]
73. Ghazi-Zahedi, K.; Langer, C.; Ay, N. Morphological computation: Synergy of body and brain. Entropy 2017, 19, 456. [CrossRef]
74. Maity, A.K.; Chaudhury, P.; Banik, S.K. Information theoretical study of cross-talk mediated signal transduction in MAPK pathways. Entropy 2017, 19, 469. [CrossRef]
75. Sootla, S.; Theis, D.; Vicente, R. Analyzing information distribution in complex systems. Entropy 2017, 19, 636. [CrossRef]
76. Tax, T.; Mediano, P.A.; Shanahan, M. The partial information decomposition of generative neural network models. Entropy 2017, 19, 474. [CrossRef]
77. Wibral, M.; Finn, C.; Wollstadt, P.; Lizier, J.T.; Priesemann, V. Quantifying Information Modification in Developing Neural Networks via Partial Information Decomposition. Entropy 2017, 19, 494. [CrossRef]
78. 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. [CrossRef]
79. Finn, C.; Lizier, J.T. Quantifying Information Modification in Cellular Automata using Pointwise Partial Information Decomposition. In Artificial Life Conference Proceedings; MIT Press: Cambridge, MA, USA, 2018, pp. 386-387.
80. Rosas, F.; Mediano, P.A.; Ugarte, M.; Jensen, H. An information-theoretic approach to self-organisation: Emergence of complex interdependencies in coupled dynamical systems. Entropy 2018, 20, 793. [CrossRef]
81. Wollstadt, P.; Lizier, J.T.; Vicente, R.; Finn, C.; Martínez-Zarzuela, M.; Mediano, P.A.; Novelli, L.; Wibral, M. IDTxl: The Information Dynamics Toolkit xl: A Python package for the efficient analysis of multivariate information dynamics in networks. J. Open Source Softw. 2019, 4, 1081. doi:10.21105/joss.01081. [CrossRef]
82. Biswas, A. Multivariate information processing characterizes fitness of a cascaded gene-transcription machinery. Chaos: Interdiscip. J. Nonlinear Sci. 2019, 29, 063108. [CrossRef]
83. James, R.G.; Emenheiser, J.; Crutchfield, J. Unique information and secret key agreement. Entropy 2019, 21, 12. [CrossRef]
84. Kolchinsky, A. A novel approach to multivariate redundancy and synergy. arXiv 2019, arXiv:1908.08642.
85. Li, M.; Han, Y.; Aburn, M.J.; Breakspear, M.; Poldrack, R.A.; Shine, J.M.; Lizier, J.T. Transitions in information processing dynamics at the whole-brain network level are driven by alterations in neural gain. PLoS Comput. Biol. 2019. [CrossRef]
86. Rosas, F.; Mediano, P.A.; Gastpar, M.; Jensen, H.J. Quantifying high-order interdependencies via multivariate extensions of the mutual information. Phys. Rev. E 2019, 100, 032305. [CrossRef] [PubMed]
87. Ince, R.A. The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. arXiv 2017, arXiv:1702.01591.
88. Lizier, J.T.; Heinzle, J.; Horstmann, A.; Haynes, J.D.; Prokopenko, M. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. J. Comput. Neurosci. 2011, 30, 85-107. [CrossRef] [PubMed]
89. Vakorin, V.A.; Krakovska, O.A.; McIntosh, A.R. Confounding effects of indirect connections on causality estimation. J. Neurosci. Methods 2009, 184, 152-160. [CrossRef]
90. Novelli, L.; Wollstadt, P.; Mediano, P.A.; Wibral, M.; Lizier, J.T. Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. Netw. Neurosci. 2019, 3, 827-847. [CrossRef]
91. Deutscher, D.; Meilijson, I.; Schuster, S.; Ruppin, E. Can single knockouts accurately single out gene functions? BMC Syst. Biol. 2008, 2, 50. [CrossRef]
92. Anastassiou, D. Computational analysis of the synergy among multiple interacting genes. Mol. Syst. Biol. 2007, 3, 83. [CrossRef]
93. White, D.; Rabago-Smith, M. Genotype-phenotype associations and human eye color. J. Hum. Genet. 2011, 56, 5-7. [CrossRef]
94. Chan, T.E.; Stumpf, M.P.; Babtie, A.C. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 2017, 5, 251-267. [CrossRef]
© 2020 by the authors. 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/).

## Erratum

1. On page 3 of the published paper between (10) and (11), Bob's expected information should be given by $H(Y)$, not $H(X)$ as is stated.
2. The sentence following (20) on page 5 of the published paper is unclear, as it could be interpreted as though it is saying that the mutual information content is non-negative. What is meant here is that the non-negativity of the mutual information does not follow trivially from the properties of the mutual information content, but rather must be proved separately.
3. The caption does not entirely describe the contents of Figure 8. Specifically, the dotted outline corresponding to $h(x, y, x)$ contains unexplained lobes. The top left lobe corresponds to the additional information associated with knowing $h(x, y)$ relative to $h(x)$, while the bottom left lobe corresponds to knowing $h(x, y, z)$ relative to $h(x, y)$. This idea is more clearly depicted for two realisations in Figure 4, which Figure 8 is supposed to generalise.

## CHAPTER 6

## CONCLUSION

This thesis presents a new framework for decomposing multivariate information and was independently derived in two distinct ways. The first approach was a top-down method of defining axioms for a measure of redundant information and was presented in Chapter 4. One of the key features of this approach was the operational interpretation provided by probability mass exclusions from Chapter 3. The second approach was a bottom-up derivation based upon the algebraic structure of shared information and was presented in Chapter 5 . One of the main results was that the redundancy lattice from partial information decomposition can be derived by combining the algebraic structures of joint and shared information content.

### 6.1 Summary of the Main Results

The partial information decomposition of Williams and Beer provided an intriguing framework for the decomposition of multivariate information [1] . However, it was not long before "serious flaws" [2, p. 2163] were identified. Despite the issues, the axiomatic derivation of the redundancy lattice seemed too elegant to be abandoned. This thesis represents a new framework for multivariate information decomposition, which was derived in two separate ways.

The first derivation was based upon William and Beer's original approach. It begins in Chapter 3, by considered the relationship between information and exclusions. Despite appearing in some of the earliest works in information theory, this description had never been formalised. The key result from this chapter is that this characterisation leads to a natural decomposition of the potentially negative pointwise mutual information content into two non-negative components-the specificity and the ambiguity. Crucially, unlike the pointwise mutual information, there is a one-to-one correspondence between the probability mass exclusions and the specificity and the ambiguity. Since both the specificity and the ambiguity are non-negative, we could then evaluate a partial information decomposition separately for each component. This yielded two separate redundancy lattices-the specificity and the ambiguity lattices. These lattices were the key result in Chapter 4.

Then based upon an operational interpretation of redundancy developed in Chapter 3, measures of pointwise redundant specificity and pointwise redundant ambiguity were defined. Together with specificity and ambiguity lattices, these measures were used to decompose multivariate information for an arbitrary number of variables. This is a key result, as many of the proposed approaches to information decomposition do not work for more than two variables. Finally, upon recombination, the resultant measure of pointwise redundant information satisfies the target chain rule. Again, this is
a key result as it means that we will get consistent results when applying this decomposition to an arbitrary number of target variables.

The second derivation is entirely independent of the first, and begins by asking the following question: if two marginal observers, Alice and Bob, share their information with a third non-observing party, Eve, such that she knows which joint realisation has occurred, and she knows the marginal probability distributions, but she does not know the joint distribution, then how much information does Eve have? We then go on to show that the algebraic structure of shared marginal information is that of a distributive lattice-that is, each distinct way in which a set of marginal observers can share their information with Eve corresponds to an element in a free distributive lattice.

By then using the fundamental theorem of distributive lattices, we showed that these new measures are isomorphic to the set union and intersections. This isomorphism is similar to Yeung's correspondence between multivariate mutual information and signed measure [3, 4]. However, in contrast to Yeung's correspondence, the measures of information content presented in this paper are non-negative. Moreover, these measures maintain a clear operational meaning regardless of the number of realisations or variables involved.

We then combine the lattice of shared marginal information content together with the semi-lattice of joint information, and show that the redundancy lattice from partial information decomposition is embedded within this larger algebraic structure. However, since we are considering marginal information contents, this structure is actually equivalent to the specificity lattice from pointwise partial information decomposition which completes the second independent derivation.

### 6.2 Further Theoretical Research

There is a need to bridge the divide between the theoretical and applied communities who utilise information theory to quantify multivariate interdependence. There is a particular need for enabling tools that would allow applied researchers to benefit from the recent theoretical advancements in information theory, such that these improvements can be applied to real-world problems. In order to achieve this, we will need to work from both the theoretical and applied sides of a particular problem such that there is an appreciation of the kinds of questions and problems that practitioners seek to answer, and an understanding of the theoretical limitations that arise when applying the framework. There are several theoretical issues which need to be addressed before information decomposition will garner widespread use.

Firstly, Williams \& Beer [1] showed that a target component can depend on two source components in four distinct ways: for three source components there are 18 distinct ways; for four components there are 166 ways; for five there are 7,579 ways; for six there are $7,828,352$ ways; for nine components the number of distinct ways is so large that, even with modern supercomputers, it is not yet computable. This scaling presents difficulties for even relatively small systems. Realistically, for most pertinent research questions, practitioners only need to quantify the most significant or important interactions. Thus, the question naturally arises: how can we quickly identify the most important interactions, and how can we represent or summarise these dependencies in an efficient and meaningful manner? Addressing this problem would enable researchers to readily quantify multivariate dependencies in non-trivial systems, and perhaps even enable the analysis of the interdependencies in complex systems.

Secondly, many real-world systems, such as bond yields or gene expression levels, are best described using continuous values. Notwithstanding this, almost all of the existing theoretical work on information decomposition considers only discrete probability. The existing theory can only be used to analyse these systems if one bins the continuous data; nevertheless, binning continuous data introduces additional parameters, undesirable artefacts, and a loss of accuracy [5]. The only existing theoretical work on the information decomposition of continuous data is limited to Gaussian systems, and so is narrow in scope [6]. Thus, there is a need to extend this framework such that it can be applied to continuous data, and hence enable a greater number of researchers to apply the tools of information decomposition to their research problems.

Thirdly, evaluating the established information-theoretic measures from continuous data will require a numerical estimator, e.g. Gaussian, box-kernel or Kraskov-Stögbauer-Grassberger [7]. These estimators each use a different algorithm which each make a distinct assumption about the underlying data. It is unclear if these numerical estimators will be suitable for the evaluation of the measures from the information decomposition. Furthermore, it is not yet clear what effect the curse of dimensionality will have upon this estimation of these measures. Addressing this issue is a necessary step for enabling practitioners to numerically quantify multivariate interdependence in continuous systems.

Finally, there is a need to provide open-source software which evaluates the decomposition in a user friendly manner. The two existing software packages for decomposing multivariate information are too limited in scope. Firstly, the Discrete Information Theory (DIT) package [8] focuses on providing theoreticians with a means for comparing their methods on hypothetical data. Secondly, while the Information Dynamics Toolkit XL (IDTxl) [9] does focus on providing methods for analysing empirical data, the technique it uses for evaluating the information decomposition is currently based upon the measure UII from Bertschinger et al. [10], which is only capable of decomposing the information provided by two source variables. Providing an implementation of the framework presented in this thesis would enable researchers to decompose multivariate information from an arbitrarily large number of source variables.

### 6.3 Potential Applications

There are certain areas of application which would be amenable to providing a proof of concept study for the use of information decomposition. For example, in the financial markets, one could consider how various factors, such as changing monetary policy or inflationary expectations, drive changes in the yields of a set of similar bonds over a range of maturities. These changes are typically highly correlated, although certain factors may lead to relative changes in the yields. In many situations, it is interesting to ask which bonds are driving these changes-information decomposition has the potential to provide quantitative answers to this question.

It is well-known that the brain uses distributed multivariate patterns to encode information about its embodied environment, and that the dynamics of individual regions in the brain depend synergistically on multiple other regions [11-13]. Information decomposition offers a means to quantify these synergistic dependencies in neural data [14], to provide a more detailed understanding of neural encoding, to improve our ability to infer brain networks, and to reveal the dynamics of how information is fused during cognitive tasks.

The vast majority of the existing analysis of phenotypic traits are based upon the pairwise relationship between the genes and a particular trait. Nevertheless, it is well-known that many traits are not dependent on any one gene, but rather are synergistically dependent on two or more genes [15, 16]; e.g. human eye colour is dependent on up to 16 different genes [17]. Information decomposition provides a means to quantify the unique, redundant and synergistic dependence between a trait and a set of genes, which promises to help us understand how phenotypes emerge from genes and hence genetic disorders.

### 6.4 References

1. Williams, P. L. \& Beer, R. D. Nonnegative decomposition of multivariate information. arXiv preprint arXiv:1004.2515 (2010).
2. Bertschinger, N., Rauh, J., Olbrich, E., Jost, J. \& Ay, N. Quantifying unique information. Entropy 16, 2161-2183 (2014).
3. Yeung, R. W. A new outlook on Shannon's information measures. IEEE Transactions on Information Theory 37, 466-474. ISSN: 0018-9448 (May 1991).
4. Yeung, R. W. Information theory and network coding (Springer Science \& Business Media, 2008).
5. Kantz, H. \& Schreiber, T. Nonlinear time series analysis (Cambridge university press, 2004).
6. Barrett, A. B. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. Physical Review E 91, 052802 (2015).
7. Lizier, J. T. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. Frontiers in Robotics and AI 1, 11 (2014).
8. James, R. G., Ellison, C. J. \& Crutchfield, J. P. dit: a Python package for discrete information theory. Journal of Open Source Software 3, 738 (2018).
9. Wollstadt, P. et al. The Information Dynamics Toolkit xl: a Python package for the efficient analysis of multivariate information dynamics in networks. Journal of Open Source Software 4, 1081 (2019).
10. 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 (2013), 251-269.
11. Lizier, J. T., Heinzle, J., Horstmann, A., Haynes, J.-D. \& Prokopenko, M. Multivariate informationtheoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. Journal of computational neuroscience 30, 85-107 (2011).
12. Vakorin, V. A., Krakovska, O. A. \& McIntosh, A. R. Confounding effects of indirect connections on causality estimation. Journal of neuroscience methods 184, 152-160 (2009).
13. Novelli, L., Wollstadt, P., Mediano, P., Wibral, M. \& Lizier, J. T. Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. arXiv preprint arXiv:1902.06828 (2019).
14. Wibral, M., Lizier, J. T. \& Priesemann, V. Bits from brains for biologically inspired computing. Frontiers in Robotics and AI 2, 5 (2015).
15. Deutscher, D., Meilijson, I., Schuster, S. \& Ruppin, E. Can single knockouts accurately single out gene functions? BMC Systems Biology 2, 50 (2008).
16. Anastassiou, D. Computational analysis of the synergy among multiple interacting genes. Molecular systems biology 3, 83 (2007).
17. White, D. \& Rabago-Smith, M. Genotype-phenotype associations and human eye color. Journal of human genetics 56, 5 (2011).

## Appendix A

## QUANTIFYING Information MODIFICATION

The conference paper here was presented at the Conference on Artificial Life 2018, in Tokyo. It contains preliminary results that demonstrate how one might use the measures of synergistic information from the framework presented in this thesis to quantify information modification in the information dynamics framework. Crucially, this new framework is uniquely suited to this application as it is the only approach which works for more than two variables and can be applied at a pointwise scale, both of which are necessary for this application.

# Quantifying Information Modification in Cellular Automata using Pointwise Partial Information Decomposition 

Conor Finn ${ }^{1,2}$ and Joseph T. Lizier ${ }^{1}$<br>${ }^{1}$ Centre for Complex Systems, The University of Sydney, NSW 2006, Australia.<br>${ }^{2}$ CSIRO Data61, Marsfield NSW 2122, Australia. conor.finn@sydney.edu.au


#### Abstract

Pointwise partial information decomposition provides a means to quantify information modification in discrete systems exhibiting intrinsic distributed computation. In his seminal "Computation at the Edge of Chaos", Chris Langton investigated how intrinsic computation emerges in cellular automata which support the three primitive functions of computation-information storage, transfer, and modification. Despite the appealing description, Langton gave no precise information-theoretic definition of the three primitive functions. In the decades since, information storage and transfer have been defined; however, a satisfactory definition of information modification has proven to be more elusive. This paper uses the recently introduced pointwise partial information decomposition to provide a quantitative measure of information modification. Moreover, this approach provides a hierarchy of different types of modifications, which each combine or synthesis different combinations of stored or transferred information. This ability to identify different types of information modification events in both space and time is exemplified with an application to cellular automata.


## Background

Understanding how distributed systems perform intrinsic computation is a central interest in the fields of artificial life, complex systems, and neuroscience. This information processing is often parsed into three fundamental components: information storage, transfer, and modification. Cellular automata, simple discrete dynamical systems from which coherent structures known as particles emerge, have long been the choice model for exhibiting distributed computation. The typical conjecture is that stationary particles store information, moving particles transfer information, and colliding particles modify information (Langton, 1990). Recently, there has been an effort to formally quantify the three component operations using information-theoretic definitions. In previous work on information dynamics, Lizier et al. $(2008,2012)$ demonstrated how information storage and transfer can be defined in terms of pointwise information measures. Crucially, this pointwise perspective enables these measures to pinpoint where and when information is being stored and transferred within the distributed system. However, this perspective has not yet delivered a satisfactory measure of information modification (Lizier et al., 2013). Here we show how pointwise
partial information decomposition (Finn and Lizier, 2018) can be used to provide a quantification of information modification which is compatible with information dynamics. We demonstrate this with an application to cellular automata.

## Overview

Information modification is interpreted to mean interactions between stored and transferred information which results in a change in this information. Using this interpretation, Lizier et al. (2013) proposed how to use the partial information decomposition (Williams and Beer, 2010) to quantify information modification. Based upon three axioms, the partial information decomposition divides the information provided a set of sources about a target into the following atoms of partial information: the information provided uniquely by each source, the information provided redundantly by two or more sources, the information provided synergistically by two or more sources, and various combinations of these three types. Lizier et al. (2013) suggested that the non-modified information in the target is any information that is identifiable in any of the sources individually; in terms of the partial information decomposition, this corresponds to the partial information atoms associated with individual sources. Conversely, the modified information is any information which is not identifiable in any of the sources individually, but is identifiable in the sources jointly; in terms of the partial information decomposition, this corresponds to all partial information atoms not accounted for previously.

Nevertheless, Lizier et al. (2013) noted two issues with their proposal. Firstly, in order to actually evaluate the partial information atoms one must define a measure of redundant information which satisfies the aforementioned axioms. There is, however, an ongoing debate as to the properties this measure of redundancy should fulfil. Many of the proposed measures can only provide a decomposition in the case of two sources. This is not sufficient for a measure of information modification since more than two information sources may be utilised in intrinsic computation. Secondly, the partial information decomposition does not provide pointwise measures of unique, redundant, and synergistic information making it incompatible with the information dynamics approach.

## Results

Recently, we took the axiomatic approach of Williams and Beer (2010) and applied it on a pointwise scale to provide measures of pointwise unique, redundant, and synergistic information (Finn and Lizier, 2018). Crucially, this pointwise partial information decomposition works for an arbitrary number of source variables and hence overcomes both of the aforementioned issues. We demonstrate how the pointwise measures of synergistic information can be used to pinpoint where and when information modification is occurring a distributed system. Moreover, since the decomposition works for an arbitrary number of information sources, it can identify a hierarchy of different orders of information modificationthe higher the order, the more information sources involved in the processing. When applied to elementary cellular automata, we get the following results: the non-modified, order one information, which is simply translated from the past or a neighbouring cell, dominates in the background domains; the modified, order two information, which involves a nontrivial synthesis of information from two sources, dominates where gliders interact with domains; finally, the modified, order three information, which combines information from all three sources, is particularly prevalent in glider collisions. We observe that Class I and II cellular automata tend to be devoid of information modification events, while Class III cellular automata are dominated by information modification events. As exemplified in Fig. 1, Class IV cellular automata feature a balance of modified and non-modified information, enabling the system to store, transfer, and modify information at different locations in space and time.

## References

Finn, C. and Lizier, J. T. (2018). Pointwise partial information decomposition using the specificity and ambiguity lattices. Entropy, 20(4).

Langton, C. G. (1990). Computation at the edge of chaos: phase transitions and emergent computation. Physica D: Nonlinear Phenomena, 42(1-3):12-37.

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

Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2008). Local information transfer as a spatiotemporal filter for complex systems. Physical Review E, 77(2):026110.

Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2012). Local measures of information storage in complex distributed computation. Information Sciences, 208:39-54.

Williams, P. L. and Beer, R. D. (2010). Nonnegative decomposition of multivariate information. arXiv preprint arXiv:1004.2515.


Figure 1: Class IV cellular automata feature a balance modified and non-modified information, enabling information storage, transfer, and modification to coexist at different locations in distributed computation. Top: elementary cellular automaton rule 54 initiated with random initial conditions. Middle top: the non-modified, order one information dominates in the background domains. Middle bottom: the modified, order two information is predominant where gliders interact with domains. Bottom: the modified, order three information modification is especially relevant in glider collisions.

## Appendix B

## Implementing the BROJA-Measure in IDTxL

Towards the beginning of my candidature, together with the co-authors of this paper, we designed and implemented a numerical algorithm for evaluating the unique information using the BROJAmeasure UI. The algorithm forms the partial information decomposition calculator in the IDTxl toolbox.

# IDTxI: The Information Dynamics Toolkit xl: a Python package for the efficient analysis of multivariate information dynamics in networks 

Patricia Wollstadt ${ }^{1}$, Joseph T. Lizier ${ }^{2}$, Raul Vicente ${ }^{3}$, Conor Finn ${ }^{2,4}$, Mario Martinez-Zarzuela ${ }^{5}$, Pedro Mediano ${ }^{6}$, Leonardo Novelli ${ }^{2}$, and Michael Wibral ${ }^{1,7,8}$<br>1 MEG Unit, Brain Imaging Center, Goethe-University Frankfurt, Fankfurt am Main, Germany 2 Centre for Complex Systems, Faculty of Engineering and IT, The University of Sydney, Sydney, Australia 3 Computational Neuroscience Lab, Institute of Computer Science, Tartu, Estonia 4 Data61, CSIRO, Epping, Australia 5 Communications and Signal Theory and Telematics Engineering, University of Valladolid, Valladolid, Spain 6 Computational Neurodynamics Group, Department of Computing, Imperial College London, London, United Kingdom 7 Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany 8 Campus Institute for Dynamics of Biological Networks, Georg-August Universität, Göttingen, Germany

## DOI: 10.21105/joss. 01081

## Software

- Review ©
- Repository ©
- Archive $\quad$ 厄

Submitted: 11 November 2018
Published: 19 February 2019

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

## Summary

We present IDTxl (the Information Dynamics Toolkit xl), a new open source Python toolbox for effective network inference from multivariate time series using information theory, available from GitHub (https://github.com/pwollstadt/IDTxl).

Information theory (Cover \& Thomas, 2006; MacKay, 2003; Shannon, 1948) is the mathematical theory of information and its transmission over communication channels. Information theory provides quantitative measures of the information content of a single random variable (entropy) and of the information shared between two variables (mutual information). The defined measures build on probability theory and solely depend on the probability distributions of the variables involved. As a consequence, the dependence between two variables can be quantified as the information shared between them, without the need to explicitly model a specific type of dependence. Hence, mutual information is a model-free measure of dependence, which makes it a popular choice for the analysis of systems other than communication channels.
Transfer entropy (TE) (Schreiber, 2000) is an extension of mutual information that measures the directed information transfer between time series of a source and a target variable. TE has become popular in many scientific disciplines to infer dependencies and whole networks from data. Notable application domains include neuroscience (Wibral, Vicente, \& Lindner, 2014) and dynamical systems analysis (Lizier, Prokopenko, \& Zomaya, 2014) (see Bossomaier, Barnett, Harré, \& Lizier (2016) for an introduction to TE and a comprehensive discussion of its application). In the majority of the applications, TE is used in a bivariate fashion, where information transfer is quantified between all sourcetarget pairs. In a multivariate setting, however, such a bivariate analysis may infer spurious or redundant interactions, where multiple sources provide the same information about the target. Conversely, bivariate analysis may also miss synergistic interactions between multiple relevant sources and the target, where these multiple sources jointly transfer more information into the target than what could be detected from examining

[^0]source contributions individually. Hence, tools for multivariate TE estimation, accounting for all relevant sources of a target, are required. An exhaustive multivariate approach is computationally intractable, even for a small number of potential sources in the data. Thus, a suitable approximate approach is needed. Although such approaches have been proposed (e.g., Lizier \& Rubinov (2012) and Faes, Nollo, \& Porta (2011)) and first software implementations exist (Montalto, Faes, \& Marinazzo, 2014), there is no current implementation that deals with the practical problems that arise in multivariate TE estimation. These problems include the control of statistical errors that arise from testing multiple potential sources in a data set, and the optimization of parameters necessary for the estimation of multivariate TE.

IDTxl provides such an implementation, controlling for false positives during the selection of relevant sources and providing methods for automatic parameter selection. To estimate multivariate TE, IDTxl utilises a greedy or iterative approach that builds sets of parent sources for each target node in the network through maximisation of a conditional mutual information criterion (Faes et al., 2011; Lizier \& Rubinov, 2012). This iterative conditioning is designed to both removes redundancies and capture synergistic interactions in building each parent set. The conditioning thus automatically constructs a non-uniform, multivariate embedding of potential sources (Faes et al., 2011) and optimizes source-target delays (Wibral et al., 2013). Rigorous statistical controls (based on comparison to null distributions from time-series surrogates) are used to gate parent selection and to provide automatic stopping conditions for the inference, requiring only a minimum of user-specified settings.

Following this greedy approach, IDTxl implements further algorithms for network inference (multivariate mutual information, bivariate mutual information, and bivariate transfer entropy), and provides measures to study the dynamics of various information flows on the inferred networks. These measures include active information storage (AIS) (Lizier, Prokopenko, \& Zomaya, 2012) for the analysis of information storage within network nodes, and partial information decomposition (PID) (Bertschinger, Rauh, Olbrich, Jost, \& Ay, 2014; Makkeh, Theis, \& Vicente, 2018; Williams \& Beer, 2010) for the analysis of synergistic, redundant, and unique information two source nodes have about one target node. Where applicable, IDTxl provides the option to return local variants of estimated measures (Lizier, 2014a). Also, tools are included for group-level analysis of the inferred networks, e.g. comparing between subjects or conditions in neural recordings.

The toolkit is highly flexible, providing various information-theoretic estimators for the user to select from; these handle both discrete and continuous time-series data, and allow choices, e.g. using linear Gaussian estimators (i.e. Granger causality, Granger (1969)) for speed versus nonlinear estimators (e.g. Kraskov, Stögbauer, \& Grassberger (2004)) for accuracy (see the IDTxl homepage for details). Further, estimator implementations for both CPU and GPU compute platforms are provided, which offer parallel computing engines for efficiency. IDTxl provides these low-level estimator choices for network analysis algorithms but also allows direct access to estimators for linear and nonlinear estimation of (conditional) mutual information, TE, and AIS for both discrete and continuous data. Furthermore low-level estimators for the estimation of PID from discrete data are provided.

The toolkit is a next-generation combination of the existing TRENTOOL (Lindner, Vicente, Priesemann, \& Wibral, 2011) and JIDT (Lizier, 2014b) toolkits, extending TRENTOOL's pairwise transfer entropy analysis to a multivariate one, and adding a wider variety of estimator types. Further, IDTxl is Python3 based and requires no proprietary libraries. The primary application area for IDTxl lies in analysing brain imaging data (import tools for common neuroscience formats, e.g. FieldTrip, are included). However, the toolkit is generic for analysing multivariate time-series data from any discipline. This is realised by providing a generic data format and the possibility to easily extend the toolkit by adding import or export routines, by adding new core estimators, or by adding

[^1]new algorithms based on existing estimators.

## References

Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., \& Ay, N. (2014). Quantifying unique information. Entropy, 16(4), 2161-2183. doi:10.3390/e16042161
Bossomaier, T., Barnett, L., Harré, M., \& Lizier, J. T. (2016). An introduction to transfer entropy. Cham, Switzerland: Springer. doi:10.1007/978-3-319-43222-9
Cover, T. M., \& Thomas, J. A. (2006). Elements of information theory. New York, NY: John Wiley \& Sons. doi:10.1002/047174882X

Faes, L., Nollo, G., \& Porta, A. (2011). Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique. Physical Review E-Statistical, Nonlinear, and Soft Matter Physics, 83, 1-15. doi:10.1103/PhysRevE. 83.051112

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3), 424-438. doi:10.2307/1912791

Kraskov, A., Stögbauer, H., \& Grassberger, P. (2004). Estimating mutual information. Physical Review E-Statistical, Nonlinear, and Soft Matter Physics, 69, 066138. doi:10. 1103/PhysRevE.69.066138
Lindner, M., Vicente, R., Priesemann, V., \& Wibral, M. (2011). TRENTOOL: A Matlab open source toolbox to analyse information flow in time series data with transfer entropy. BMC Neuroscience, 12(1), 119. doi:10.1186/1471-2202-12-119
Lizier, J. T. (2014a). Measuring the dynamics of information processing on a local scale in time and space. In M. Wibral, R. Vicente, \& J. T. Lizier (Eds.), Directed information measures in neuroscience, Understanding complex systems (pp. 161-193). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-54474-3_7

Lizier, J. T. (2014b). JIDT: An Information-Theoretic Toolkit for Studying the Dynamics of Complex Systems. Frontiers in Robotics and AI, 1(December), 11. doi:10.3389/frobt. 2014.00011

Lizier, J. T., \& Rubinov, M. (2012). Multivariate construction of effective computational networks from observational data. Max Planck Institute: Preprint, Preprint no. 25.
Lizier, J. T., Prokopenko, M., \& Zomaya, A. Y. (2012). Local measures of information storage in complex distributed computation. Information Sciences, 208, 39-54. doi:10. 1016/j.ins.2012.04.016

Lizier, J. T., Prokopenko, M., \& Zomaya, A. Y. (2014). A framework for the local information dynamics of distributed computation in complex systems. In M. Prokopenko (Ed.), Guided self-organization: Inception, Emergence, complexity and computation (Vol. 9, pp. 115-158). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-53734-9_5
MacKay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge: Cambridge University Press.

Makkeh, A., Theis, D. O., \& Vicente, R. (2018). BROJA-2PID: A robust estimator for bertschinger et al.'s bivariate partial information decomposition. Entropy, 20(4), 271. doi:10.3390/e20040271

Montalto, A., Faes, L., \& Marinazzo, D. (2014). MuTE : a MATLAB toolbox to compare established and novel estimators of the multivariate transfer entropy. PLoS ONE, 9(10), e109462. doi:10.1371/journal.pone. 0109462

[^2]Schreiber, T. (2000). Measuring information transfer. Physical Review Letters, 85(2), 461-464. doi:10.1103/PhysRevLett.85.461
Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379-423. doi:10.1145/584091.584093

Wibral, M., Pampu, N., Priesemann, V., Siebenhühner, F., Seiwert, H., Lindner, M., Lizier, J. T., et al. (2013). Measuring information-transfer delays. PloS ONE, 8(2), e55809. doi:10.1371/journal.pone.0055809
Wibral, M., Vicente, R., \& Lindner, M. (2014). Transfer entropy in neuroscience. In M. Wibral, R. Vicente, \& J. T. Lizier (Eds.), Directed information measures in neuroscience (pp. 3-36). Springer. doi:10.1007/978-3-642-54474-3_1
Williams, P. L., \& Beer, R. D. (2010). Nonnegative decomposition of multivariate information.

## Appendix C

## An Application in Neuroscience

The partial information decomposition from the paper in Appendix B was utilised in this paper.

Article

# Quantifying Information Modification in Developing Neural Networks via Partial Information Decomposition 

Michael Wibral ${ }^{1,2, *}$ (iD) , Conor Finn ${ }^{3,4}$ (iD) Patricia Wollstadt ${ }^{1}$, Joseph T. Lizier ${ }^{3}$ (iD) and Viola Priesemann ${ }^{2,5}$<br>1 MEG Unit, Brain Imaging Center, Goethe University, 60528 Frankfurt, Germany; patricia.wollstadt@gmx.de<br>2 Max Planck Institute for Dynamics and Self-Organization, 37077 Göttingen, Germany; vp@ds.mpg.de<br>3 Complex Systems Research Group and Centre for Complex Systems, Faculty of Engineering \& IT, The University of Sydney, Sydney, NSW 2006, Australia; conor.finn@sydney.edu.au (C.F.); joseph.lizier@sydney.edu.au (J.T.L.)<br>4 CSIRO Data61, Marsfield, NSW 2122, Australia<br>5 Bernstein Center for Computational Neuroscience, 37077 Göttingen, Germany<br>* Correspondence: wibral@em.uni-frankfurt.de

Received: 13 July 2017; Accepted: 12 September 2017; Published: 14 September 2017


#### Abstract

Information processing performed by any system can be conceptually decomposed into the transfer, storage and modification of information-an idea dating all the way back to the work of Alan Turing. However, formal information theoretic definitions until very recently were only available for information transfer and storage, not for modification. This has changed with the extension of Shannon information theory via the decomposition of the mutual information between inputs to and the output of a process into unique, shared and synergistic contributions from the inputs, called a partial information decomposition (PID). The synergistic contribution in particular has been identified as the basis for a definition of information modification. We here review the requirements for a functional definition of information modification in neuroscience, and apply a recently proposed measure of information modification to investigate the developmental trajectory of information modification in a culture of neurons vitro, using partial information decomposition. We found that modification rose with maturation, but ultimately collapsed when redundant information among neurons took over. This indicates that this particular developing neural system initially developed intricate processing capabilities, but ultimately displayed information processing that was highly similar across neurons, possibly due to a lack of external inputs. We close by pointing out the enormous promise PID and the analysis of information modification hold for the understanding of neural systems.


Keywords: information theory; partial information decomposition; neural computation; neural development; self-organisation

## 1. Introduction

Shannon's quantitative description of information and its transmission through a communication channel via the entropy and the channel capacity, respectively, has drawn considerable interest from the field of neuroscience from the very beginning. This is because information processing in neural systems is typically performed in a highly distributed way by many communicating processing elements, the neurons.

However, in contrast to a channel in Shannon's sense, the purpose of dendritic connections in a neural system is not to simply relay information for the sake of reliable communication. Instead, communication between neurons serves the purpose of collecting multiple streams of information at a neural processing element that modifies this information, i.e., that synthesizes the
incoming streams into output information that is not available from any of these streams in isolation. This becomes immediately clear when looking at the meshed structure of nervous systems, where multiple communication streams converge on single neurons, and where neural output signals are sent in a divergent manner to many different receiving neurons. This structure differs dramatically from a structure solely focused on the reliable transmission of information where many parallel, but non-interacting streams would suffice. Thus, the meshed architecture seems to have evolved to "fuse" information from different input sources (including a neuron's recent spiking history and its current state) in a nontrivial way, e.g., other than simply multiplexing it in the output. In other words, the distributed computation in neural systems may heavily rely on information modification [1].

Attempts at formally defining information modification have presented a considerable challenge, however, in contrast to the well established measures of information transfer [2-6] and active information storage [7-9]. This is because identifying the "modified" information in the output of a processing element amounts to distinguishing it from the information from any input that survives the passage through the processor in unmodified form. These unmodified parts, in turn, may either come uniquely from one of the inputs, uniquely from another input (unique mutual information between an input and the output), or may be provided by multiple inputs simultaneously (shared mutual information between several inputs and the output). A decomposition of the mutual information between the inputs and the output of this kind is called a partial information decomposition (PID) [10] (Some authors prefer the simpler term "information decomposition", also in this special issue.).

In a PID, the problem of identifying modified information is equivalent to identifying the part of the (joint) mutual information that is not unique mutual information from one or another input, and that is also not shared mutual information from multiple inputs. This remaining part has been termed synergistic mutual information in the work of Williams and Beer [10], and has been identified with information modification in [11] for the reasons given above.

The recent pioneering work by Williams and Beer [10] revealed that the standard axioms of information theory do not uniquely define the unique, shared and synergistic contributions to the mutual information, and that additional axioms must be chosen for its meaningful decomposition. Among several possible choices of additional axioms or assumptions available at the time of this study (see Section 4.1) we here adhere to the definition given independently by Bertschinger, Rauh, Olbrich, Jost, and Ay ("BROJA-measure", [12]) and Griffith and Koch [13]. Our decision is based on two properties of the BROJA measure that seem necessary for an application to the problem of information modification as described above: first, in their definition, the presence of non-zero synergistic mutual information for the case of two inputs and one output cannot be deduced from the (two) marginal distributions of one input and one output variable. This property distinguishes the BROJA measure from the others available at the time; Bertschinger et al. [12] referred to it as Assumption (**), and showed that the BROJA measure is the only measure that satisfies both Assumption ( $*$ ) and $(* *)$, where Assumption $(*)$ indicates that the existence of unique information only depends on the pairwise marginal distributions between the individual inputs and the output. The measures from Williams and Beer [10] and Harder et al. [14] only satisfy Assumption (*), but not (**)—see [12] for details. Second, the BROJA measure is placed on a rigorous mathematical footing, being derived directly from the aforementioned Assumption $(*)$ rather than postulated ad hoc; furthermore, it has an operational interpretation in terms of expected utilities from the output based on knowledge of only each input, and many mathematical properties proven.

In our proof-of-principle study, we apply the BROJA decomposition of mutual information to the analysis of the emergent information processing in self-organizing neural cultures, and show that these novel information theoretic concepts indeed provide a meaningful contribution to our understanding of neural computation in this system.

## 2. Methods

In the following, we consider the neural data produced by two neurons as coming from two stationary random processes $\mathcal{X}_{1}, \mathcal{X}_{2}$, composed of random variables $X_{1}(i)$ and $X_{2}(i), i=1 \ldots n$, with realizations $x_{1}(i), x_{2}(i)$. The corresponding embedding or state space vectors are given in bold font, e.g., $\mathbf{x}^{\mathbf{1}}(\mathbf{i})=\{x(i), x(i-1) \ldots, x(i-l+1)\}$. The state space vector $\mathbf{x}^{\mathbf{1}}(\mathbf{i})$ is constructed such that it renders the variable $x(i+1)$ conditionally independent of all random variables $x(j)$ with $j<i-l+1$, i.e., $p\left(x(i+1) \mid \mathbf{x}^{\mathbf{1}}(\mathbf{i}), x(j)\right)=p\left(x(i+1) \mid \mathbf{x}^{\mathbf{1}}(\mathbf{i})\right)$.

### 2.1. Definition and Estimation of Unique, Shared and Synergistic Mutual Information

For two input random variables $X_{1}, X_{2}$ and an output random variable $Y$ (Figure 1), Bertschinger et al. [12] defined the four unique, shared and synergistic contributions to the joint mutual information $I\left(Y: X_{1}, X_{2}\right)$ as:

$$
\begin{align*}
\tilde{I}_{u n q}\left(Y: X_{1} \backslash X_{2}\right): & =\min _{Q \in \Delta_{P}} I_{Q}\left(Y: X_{1} \mid X_{2}\right)  \tag{1}\\
\tilde{I}_{u n q}\left(Y: X_{2} \backslash X_{1}\right): & =\min _{Q \in \Delta_{P}} I_{Q}\left(Y: X_{2} \mid X_{1}\right)  \tag{2}\\
\tilde{I}_{\text {shd }}\left(Y: X_{1} ; X_{2}\right): & =\max _{Q \in \Delta_{P}}\left(I\left(Y: X_{1}\right)-I\left(Y: X_{1} \mid X_{2}\right)\right)  \tag{3}\\
\tilde{I}_{\text {syn }}\left(Y: X_{1} ; X_{2}\right): & =I\left(Y:\left(X_{1}, X_{2}\right)\right) \\
& -\min _{Q \in \Delta_{P}} I_{Q}\left(Y:\left(X_{1}, X_{2}\right)\right) \tag{4}
\end{align*}
$$

where $I$ is the standard mutual or conditional mutual information $[15,16], \tilde{I}_{u n q}$ is the unique, $\tilde{I}_{\text {shd }}$ the shared, and $\tilde{I}_{s y n}$ the synergistic mutual information. In our notation, the comma separates variables within a set that are considered jointly, the colon separates the (sets of) random variables between which the mutual information is computed, while the semicolon or backslash separates sets of random variables that we are decomposing such mutual information across. For the latter, the semicolon is used for measures where the sets of random variables are considered symmetrically (i.e., shared and synergistic information), while the backslash is used for asymmetric cases (i.e., unique information in one but not the other). $\Delta_{Q}$ in the above definitions is the space of probability distributions $Q$ that have the same pairwise marginal distributions between each input and the output as the original joint distribution $P$ of $X_{1}, X_{2}, Y$, i.e.:

$$
\begin{array}{r}
\Delta_{P}=\left\{Q \in \Delta: Q\left(X_{1}=x_{1}, Y=y\right)=P\left(X_{1}=x_{1}, Y=y\right)\right. \\
\left.\quad \text { and } Q\left(X_{2}=x_{2}, Y=y\right)=P\left(X_{2}=x_{2}, Y=y\right)\right\} . \tag{5}
\end{array}
$$



Figure 1. Decomposition of the joint mutual information between two input variables $X_{1}, X_{2}$ and the output variable $Y$. Modified from [17], CC-BY license.

### 2.2. Mapping of Neural Recordings to Input and Output Variables for PID, and Definition of Information Modification

In our application to developing neural cultures, we always consider two spike trains $(A, B)$ at a time: the past state of the spike train $\mathrm{A}, \mathbf{X}_{\mathbf{A}}^{-}$, is one of the input variables and the past state of a spike train $B, X_{B}^{-}$, is considered as the other input variable. Empirically, these states are usually constructed using the aforementioned embedding or state space vectors $\mathbf{x}^{1}(\mathbf{i})$ of length $l$. The output variable $X_{A}^{+}$ is simply spike train A's current spiking behavior (spiking or not).

This output variable is computed from external inputs $\left(\mathbf{X}_{\mathbf{B}}^{-}\right)$as well as the output variable's own history $\mathbf{X}_{\mathbf{A}}^{-}$. When analyzing this computation, one wishes to focus on the operations of information storage, transfer and modification, in alignment with established views of distributed information processing in complex systems $[18,19]$. In this study, specifically, we will focus on information modification, yet we first need to decompose the output variable in terms of information storage and information transfer, where the latter will also contain the information modification (see [11] and Figure 2):

$$
\begin{equation*}
I\left(X_{A}^{+}: \mathbf{X}_{\mathbf{A}}^{-}, \mathbf{X}_{\mathbf{B}}^{-}\right)=I\left(X_{A}^{+}: \mathbf{X}_{\mathbf{A}}^{-}\right)+I\left(X_{A}^{+}: \mathbf{X}_{\mathbf{B}}^{-} \mid \mathbf{X}_{\mathbf{A}}^{-}\right) \tag{6}
\end{equation*}
$$

Here, $I\left(X_{A}^{+}: \mathbf{X}_{\mathbf{A}}^{-}\right)$is the active information storage [7], the predictive information from the past state of the variable to its next value. Then, $I\left(X_{A}^{+}: \mathbf{X}_{\mathbf{B}}^{-} \mid \mathbf{X}_{\mathbf{A}}^{-}\right)$is the transfer entropy [2], the predictive information from the past of the other source $B$ to the next value of $A$, in the context of the past of $A$.

In order to identify information modification, we need to take this decomposition further to reveal two sub-components of each of these information storage and transfer terms. These sub-components result from a partial information decomposition of $I\left(X_{A}^{+}: \mathbf{X}_{\mathbf{A}}^{-}, \mathbf{X}_{\mathbf{B}}^{-}\right)$into four parts (see Figure 2):

1. The unique mutual information of the output spike train's own past $\tilde{I}_{u n q}\left(X_{A}^{+}: \mathbf{X}_{\mathbf{A}}^{-} \backslash \mathbf{X}_{\mathbf{B}}^{-}\right)$—this can be considered as information uniquely stored in the past output of the spike train that reappears at the present sample.
2. The unique information from the other spike train $\tilde{I}_{u n q}\left(X_{A}^{+}: \mathbf{X}_{\mathbf{B}}^{-} \backslash \mathbf{X}_{\mathbf{A}}^{-}\right)$—this is the information that is transferred unmodified from the input to the output of the receiving spike train (also known as the state independent transfer entropy [20]).
3. The shared mutual information about the output of spike train A that can be obtained both from the past states of spike train A and of spike train B, $\tilde{I}_{s h d}\left(X_{A}^{+}: \mathbf{X}_{\mathbf{B}}^{-} ; \mathbf{X}_{\mathbf{A}}^{-}\right)$—this is information that is redundantly stored in the past of both spike trains and that reappears at the present sample.
4. The synergistic mutual information $\tilde{I}_{s y n}\left(X_{A}^{+}: \mathbf{X}_{\mathbf{B}}^{-} ; \mathbf{X}_{\mathbf{A}}^{-}\right)$, i.e., the information in the output of spike train $\mathrm{A}, \mathrm{X}_{A}^{+}$, that can only be obtained when having knowledge about both the past state of the external input, $\mathbf{X}_{\mathbf{B}}^{-}$, and the past state of the receiving spike train, $\mathbf{X}_{\mathbf{A}}^{-}$. (This is also known as the state dependent transfer entropy [20]).


Figure 2. Mapping between the decomposition into storage and transfer (A) and individual or joint mutual information terms, and PID components (B). Numbers in (B) refer to the enumeration of components given in Section 2.2. Number " 4 " is the modified information.

We see that Components 1 and 3 above form the active information storage in Equation (6), while Components 2 and 4 form the transfer entropy term. Component 4 , as the synergistic mutual information contributed by the storage and the transfer source, is what we consider to be the modified information (following [11]). The same underlying definition of information modification from [11] was used by Timme et al. [21] in an earlier study of dynamics of spiking activity of neural cultures, yet with another PID measure and considering multiple external inputs to a neuron (see discussion for further details).

### 2.3. PID Estimation

PID terms were estimated by minimizing the conditional mutual information as indicated in the first equation of the system 1. To perform the minimization, we used a stochastic approach where alternative trial distributions in $\Delta_{P}$ are created by swapping probability mass $\delta_{p}$ between the symbols of the current distribution such that the constraints defining $\Delta_{P}$ are satisfied. If this swap of probability mass leads to a reduction in the conditional mutual information, the trial distribution is made the current distribution, and a new trial distribution is created. If the trial distribution fails to reduce the conditional mutual information, then a new trial distribution is created from the current distribution. This latter process is repeated for a maximum of $n$ unsuccessful swaps in a row (here $n=20,000$ ), with a reset of the counter in case of a successful swap. If after these trials no reduction is reached, then we assume that we have found the optimum possible with the current increment in probability mass $\delta_{p}$ and that a finer resolution is needed. Hence, the increment is halved: $\delta_{p} \leftarrow \delta_{p} / 2$. This process starts with an initial $\delta_{p}$ equal to the largest probability mass assigned to any symbol in the distribution $P$, and is repeated until the numerical precision of the machine or programming language is exhausted (here, we performed 63 divisions of the original $\delta_{p}$ by a factor of 2 , using Numpy 1.11.2 under Python 3.4.3 and 128-bit floating point numbers). The algorithm is available in the open source toolbox IDT $^{x l}$ [22]. We note that better solutions based on convex optimization exist (see Makkeh et al. [23] in this special issue) and that these are implemented in newer versions of IDT ${ }^{x l}$; at the time of performing this study, however, these implementations were not available to us yet.

### 2.4. Statistical Testing

Results obtained for the joint mutual information, and for the four PID measures normalized by the joint mutual information, were subjected to pairwise statistical tests for differences in the median between recordings days (see Section 2.5) by means of permutation tests. An uncorrected $p$-value of $p<5 \times 10^{-4}$, corresponding to $p$-value of $p<0.05$ with Bonferroni correction for multiple comparisons across five measures, and 20 pairs of recording days, was considered significant.

We normalized the PID values to remove influences from changes in the overall activity of the culture (that change the entropy of the inputs) and to abstract from changes in the overall joint mutual information. Note that we did not test these normalized PID values for significance against surrogate data, as the focus here was on changes with development of the culture. Moreover, the four normalized PID terms analyzed here are not independent from each other, but instead sum up to a value of 1, making the construction of a meaningful statistical test difficult.

### 2.5. Electrophysiological Data—Acquisition and Preprocessing

The spike recordings were obtained by Wagenaar et al. [24] from a single in vitro culture of $M \approx 50,000$ cortical neurons. The data are available online at [25]; of the data provided in this repository, we used culture/experiment " $2-1$ ", days $7,14,21,28$, and 34 . Details on the preparation, maintenance and recording setting can be found in the original publication. In brief, cultures were prepared from embryonic E18 rat cortical tissue. Recordings lasted more than 30 min . The recording system comprised an $8 \times 8$ array of 59 titanium nitride electrodes with $30 \mu \mathrm{~m}$ diameter and $200 \mu \mathrm{~m}$ inter-electrode spacing, manufactured by Multichannel Systems (Reutlingen, Germany). As described in the original publication, spikes were detected online using a threshold based detector as upward
or downward excursions beyond 4.5 times the estimated root mean squared (RMS) noise [26]. Spike waveforms were stored, and used to remove duplicate detections of multiphasic spikes. Spike sorting was not employed, and thus spike data represent multi-unit activity. To obtain a tractable amount of data, we randomly picked spike time series from the dataset, and of these only selected those that developed at least a moderate level of activity with maturation of the culture (channels $01,02,03,04$, $05,07,11,13,14,16,19,50,53,57,58,60$ ), to guarantee a certain level of (Shannon) information to be present. In total, our analyses comprise 16 spike time series, i.e., 240 pairs of spike time series.

From these spike time series, the realizations $x_{A}^{+}(i)$ of the random variable $X_{A}^{+}(i)$ were constructed by applying bins of 8 ms length; empty bins were denoted by zeros, whereas bins that contained at least one spike were denoted with ones. The corresponding approximate past state vectors $\mathbf{x}_{\mathbf{A} / \mathbf{B}}^{-1}(\mathbf{i})$ were constructed with finite past length $l=3$, and to balance the need for low dimensionality for an unbiased estimate and a coverage of as much past history as possible, three past bins of size 8 ms , 32 ms and 32 ms were defined, where the shortest bin was the bin closest to $i$, and where both the 8 ms and the two 32 ms bins were set to one or zero depending on whether or not a spike occurred anywhere within these bins. This approach to cover a longer history at a low dimensionality amounts to a compressing of the information in the history of the process, aiming to retain what we perceive to be the most relevant information. This approach is similar to the one used by Timme et al. [27], except for the use of nonuniform binwidths in our case. Alternative approaches to large bin widths exist that are either based (i) on nonuniform embedding, picking the most informative past samples (or bins with a small width on the order of the inverse sampling rate) from a collection of candidates (e.g., [28-30]), and the IDT $^{x l}$ toolbox [22]; or (ii) on varying the lag between an a vector of evenly spaced past bins and the current sample [4,31,32], but both of these approaches might be less suitable for relatively sparse binary data, such as spike trains.

## 3. Results

### 3.1. PID of Information Processing in Neural Cultures

From the original report by Wagenaar et al. [26], the following aspects of the development of the "dense" culture analyzed here can be observed: (i) by preparation, neurons were unconnected and mostly silent at first; then (ii) show spontaneous activity and begin to be connected (compare the increase of mutual information between inputs and outputs in Figure 3); later, they (iii) become densely connected and thereby strongly responsive to each other's spiking activity (Quote from [26]: "We ... found that functional projections grew rapidly during the first week in vitro in dense cultures, reaching across the entire array within 15 days (Figure 9)"); while, in a last stage, (iv) connectivity often leads to activity pattern where all neurons become simultaneously active in large, culture spanning bursts of activity. This can, for example, be seen for data used here in the development of large, system spanning neural avalanches with maturation of the culture (see Figure 4 in [33] and Figure 13 in the corresponding preprint [34]; for the definition of neural avalanches as used here, see [35]). The number of such system spanning avalanches was $[0,0,7,50,73]$ for the five recording weeks. At the same time, the mean avalanche sizes (defined in [35]) also increased as (1.05, 1.31, 1.81, 4.39, 3.42)—note the jump from week 2 to 3 in both measures, and compare to the normalized shared information in Figure 4.

From the viewpoint of partial information decomposition, we hypothesized that stages (i) and (ii) should be characterized by a high fraction of unique information from a neuron's own history because neurons that do not yet receive sufficient input to trigger their firing can only have unique mutual information with their own history.

Unique information from other neuron's inputs, and also synergy between both neurons' past states should be visible in stage (iii) because we assume that neurons, even in vitro are wired to fuse information from multiple sources with their the information of their own state. Thus, we expected non-trivial computation in the form of synergy to be visible as long as the input distributions are sufficiently different from a neuron's own history.


Figure 3. Left: development of the joint mutual information with network maturation. Grey symbols and lines-joint mutual information (MI) from individual pairs of spike time series, red symbols-median over all pairs. Horizontal black lines connect significantly different pairs of median values ( $p<0.05$, permutation test, Bonferroni corrected for multiple comparisons); Right: magnification of the joint mutual information estimates in the first two recording weeks. Note that the three large outliers from week 2 have been omitted from the plot. These tiny, but non-zero, values form the basis for the normalized non-zero PID terms presented in Figure 4-also leading to considerable variance there.


Figure 4. Development of normalized PID contributions (i.e., PID terms normalized by the joint mutual information) with network maturation. Grey symbols and lines-PID values from individual pairs of spike time series, red symbols-median over all pairs. Horizontal black lines connect significantly different pairs of median values ( $p<0.05$, permutation test, Bonferroni corrected for multiple comparisons). On the lower right, note the sudden increase in normalized shared mutual information from week 2 to 3 that coincides with the onset of system spanning neural avalanches (see text).

In the last stage (iv), partial information decomposition should then be dominated by shared mutual information because when all input distributions are more or less identical and highly correlated, then there can only be shared information (at least when using the BROJA measure).

From preliminary investigations [36], we also expected the joint mutual information between both inputs and the output to rise. Given the caveat that we analyzed multi-unit activity here, instead of single units (i.e., single neurons) obtained by spike sorting, our results comply with these hypotheses: the initial two recording weeks were dominated by unique information from a spike time series' own history, while, in intermediate recording weeks, synergistic and shared information were dominant, and shared information finally prevailed in the last two recordings (Figure 3 shows the joint
mutual information, with the normalized PID contributions to this shown in Figure 4 and raw PID contributions in Figure 5).


Figure 5. Development of raw PID contributions with network maturation. Grey symbols and lines-PID values from individual pairs of spike time series, blue symbols-median over all pairs. Note that we do not provide statistical tests here as the visible differences are heavily influenced by the corresponding differences in the joint mutual information (see Figure 3).

## 4. Discussion

We here applied PID to neural spike recordings with the objective to compute a measure of information modification, and, for the first time, to assess its face validity given what is already known about information processing in developing neural cultures. Our analyses of the synergistic part of the mutual information between information storage and transfer sources, which we see as a promising candidate measure of information modification, complied with our intuition on how information modification should rise with development as neurons get connected and their synaptic weights adapt to the environment of the culture (i.e., with a lack of external input to the culture). The end of this rise in (relative) information modification and a final drop caused by a jump in the (relative) shared part in the mutual information was also expected given that a computation must always be understood as the composition of a mechanism and an input distribution. This input distribution is well known to get more and more similar over neurons as the culture approaches the typical bursting behavior that synchronizes all activity. With all input distributions being similar in this way, there is reduced scope for modifying information-hence the observed drop in the last recording week. In summary, the partial information decomposition used here and the results for its synergistic part capture well our intuition of what should happen in this simple neural system. This increases confidence in the usefulness and interpretability of PID measures in the analysis of neural data from more complex neural systems.

Two additional aspects seem important here: first, our analyses underline one of the key theoretical advances of PID, that all four PID terms, and especially shared and synergistic ones, can coexist simultaneously-a fact overlooked in early attempts to define shared (or 'redundant') and synergistic contributions to the mutual information (see references in [10]); second, no knowledge on the typical development of neural cultures was necessary to arrive at our PID results; in other words, the development of computation in the culture could have been derived from our PID analysis alone.

This makes PID of neural activity a useful first step when investigating the computational architecture of a neural system.

In the sections below, we discuss some caveats to consider with this relatively young analysis technique, where several competing definitions of a PID coexist, not all of them equally suitable for computing information modification [11]. We also expand on the aforementioned relation between measures of information transfer and modification. Moreover, we would like to highlight and expand on the fact that a computation is a composition of an input distribution and a mechanism working on these inputs. Neglecting the importance of the input distribution and understanding a PID as directly describing a computational mechanism is a frequent misunderstanding that we would like to clarify here. We close by highlighting potential uses of PID in neuroscience.

### 4.1. Which Definition of Synergistic Mutual Information to Use?

In contrast to earlier studies of synergy or information modification in neural data [21,37], we here used the definition of unique, shared and synergistic mutual information as given by Bertschinger et al. [12] and by Griffith and Koch [13] (BROJA-decomposition). As initially outlined, in our view, this definition was the only published one at the time of experiment that had the properties necessary for a mapping between information modification and the synergistic part of the mutual information, and is sufficiently easy to compute because of the convexity of the underlying optimization.

However, the BROJA definition has also been criticized because a decomposition is only possible for the case of two inputs (although these inputs themselves can be arbitrarily large groups of variables). We consider this an acceptable restriction for some purposes in neuroscience as it seems to map well to the properties of cortical neurons; for example, the pyramidal neurons of the neocortex keep exactly two classes of inputs separate via their apical and basal dendrites (see [17,38,39] and references therein). In addition, as long as one is only interested in the computations performed by a neuron based on its own history and all its inputs considered together as one (vector-valued) input variable, this framework is sufficient (see, for example, the treatment of this case in the theoretical study presented in [17]).

In contrast to the BROJA-decomposition, Williams and Beer suggested in their original work [10] an alternative definition that allowed the decomposition of the mutual information between multiple inputs (considered separately, not as a group) and an output into a partial order (a mathematical "lattice") of shared information terms. While this decomposition into a lattice of terms clearly is desirable, the measure of shared information given by Williams and Beer [10] (known as $I_{\text {min }}$ ) also has several properties that have been questioned. First, it does not respect the locality of information, i.e., point-wise interpretations of this shared information are not continuous with respect to the underlying probability distribution functions [11]. Second, it suggests the presence of shared information in situations where in each realization only a single source ever holds non-zero information about a target [40]. We note that the latter is an issue for the BROJA-decomposition as well.

Third, several authors have questioned the presence of non-zero shared mutual information under $I_{\min }$ when there is no pairwise mutual information between the inputs themselves while the output is a simple collection of these inputs (known as "two bit copy"). A desire for zero shared mutual information in this case was formalized in a so-called identity axiom by Harder et al. [14]. This axiom suggests that if two inputs $X_{1}, X_{2}$ with no mutual information between them $\left(I\left(X_{1}: X_{2}\right)=0\right)$ are combined into an output that is simply their collection, i.e., $Y=\left\{X_{1}, X_{2}\right\}$ then the shared part of the joint mutual information $I\left(Y: X_{1}, X_{2}\right)$ must be zero. However there are significant arguments against the inclusion of such an axiom, and in support of the presence of shared information in the two bit copy problem; see, e.g., Bertschinger et al. [41], and, in this issue, by Ince [42] and Finn et al. [40]. For example, there can be no measure of redundant information that simultaneously satisfies the original three PID axioms, has non-negative PI atoms, and possesses the identity property [43].

Debates continue on this aspect, and, in the future, it will be interesting to check the consistency of results reported here with respect to alternative decompositions, such as those presented by Finn et al. [40] or Ince [42] in this issue.

In summary, the BROJA measure used in this study has several appealing properties, yet it lacks the ability to decompose the information of more than two input variables into a lattice. Several contributions to this special issue present progress on lattice-compatible distributions [40,42] and also investigate the consequences of the symmetrical, or asymmetrical treatment of information sources and targets [44] (also see the work of Olbrich et al. [45] on this topic).

### 4.2. Previous Studies of Information Modification in Neural Data

Timme et al. [21] studied information modification in the dynamics of spiking activity of neural cultures with a focus on the relation between information modification at a neuron and its position in the underling (effective) network structure. They report, for example, that neurons which modify "large amounts of information tended to receive connections from high out-degree neurons". Both their study and ours have in common the same underlying definition of information modification [11]. Their study differs slightly from ours in examining synergy between two external inputs to a neuron, conditional on that neuron's past, whereas we examine synergy between one external input and the receiving neuron's past. A more important difference between their study and ours, however, is the choice of PID measure (see above). Specifically, they used the Williams and Beer [10] $I_{\text {min }}$ measure, in contrast to the BROJA measure used here-see Section 4.1 for details on the consequences of these choices.

Another important difference is the use of multi unit activity in our study, while Timme et al. [21] used spike sorted data that represents the activity of single neurons. However, for the data-set we used, spike extraction was relatively conservative, using a high threshold and removing events with spurious waveforms [26]. This resulted in a relatively low average multi-unit activity of less than 3.5 Hz . This is comparable to the mean rate of 2.1 Hz reported by Timme et al. [21]. From this, we estimate that only one or two close by neurons typically contribute to the recorded multi-unit activity. Thus, this difference may be relatively minor in practice. Conceptually, however, the information contained in single and multi-unit activity clearly differs in interpretation-see the next section for the more details.

We also note that there are earlier applications of the concept of synergy (meant as synergistic mutual information) to neural data (e.g., [46-49]) that relied on the computation of interaction information. However, when interpreting these studies, it should be kept in mind that these report the difference between shared information and synergistic information-as detailed by Williams and Beer [10]. If both are present in the data (a possibility that may simply have been overlooked by most researchers before Williams and Beer [10]), then this view of a 'net-synergy' only gives a partial view of the coding principles involved.

### 4.3. Information Represented by Multi and Single Unit Data

As detailed in the methods section, we performed our analyses on multi unit activity, i.e., we considered all spiking activity picked up by a recording electrode-potentially coming from multiple neurons. Thus, the information processing analyzed here is that of a cluster of neurons close to the recording electrodes, but not that of individual neurons, limiting the direct interpretation of our results. This problem can be alleviated by using spike sorting algorithms, e.g., based on the individual waveforms to assign each spike to an individual neuron, and then analyzing only the spikes of individual neurons. This has indeed been done in the study by Timme et al. [21] and improves the interpretation of the results in terms of neural coding. Ideally, it should be included in follow-up studies on information modification via PID as well. However, as the multi-unit activity reported here most likely contained only one or two single units (see previous section), we expect very similar results for an analysis of single units.

### 4.4. Measured Information Modification versus the Capacity of a Mechanism for Modifying Information

To appreciate the findings of the current study, it is important to realize that any computation is a composition of (i) a mechanism and (ii) an input distribution. As an extreme example, take an "exclusive or" (XOR)-gate, which has only one bit of synergy when fed by two uniformly distributed random bit inputs. However, when we clamp one of these inputs, for example $X_{1}$, to producing just ' 0 s, then all the information (still one bit of joint mutual information) is unique information from the other input $X_{2}$. This result must hold for all PID measures by virtue of the equations linking classic mutual information terms and PID terms (Equations (1)-(3) in [12], also consult Figure 1), and due to the fact that the mutual information of the clamped input and the output must be equal to or smaller than the entropy of that input, which is zero. Feeding the XOR gate with an alternative input distribution $p_{a}\left(x_{1}, x_{2}\right)$ of the form $p_{a}(0,1)=3 / 8, p_{a}(1,1)=3 / 8, p_{a}(0,0)=p_{a}(1,0)=1 / 8$ yields 0.811 bits of synergy and 0.188 of unique information from $x_{1}$, using the BROJA PID.

Another simple example would be a logical conjunction (AND)-gate fed by two different input distributions: when fed by two independent streams of input bits with uniform probabilities of zeros and ones, the BROJA PID results in 0.5 bit of synergy and 0.311 bit of shared information [12]. Feeding the same mechanism with an alternative input distribution $p_{b}\left(x_{1}, x_{2}\right)$ of the following form: $p_{b}(0,0)=3 / 8, p_{b}(1,1)=3 / 8, p_{b}(0,1)=p_{b}(1,0)=1 / 8$ results in approximately 0.406 bit of synergy and 0.549 bit of shared information as measured by the BROJA PID.

This dependence of the PID on the input distribution means that describing a computation in terms of information modification via the synergistic information describes the joint operation of input distribution and mechanism (with the consequences related to bursting activity in neural cultures that were noted above). Indeed, this is the correct information theoretic description of how the system modified information in the specific computation reflected in the data. This description does not, however, inform us about how much information modification the mechanism performing the computation is capable of in principle. This is analogous to the situation of a communication channel in Shannon's theory where the mutual information $I_{P_{X}}(X: Y)$ between the input $X$ and the output $Y$ of a channel informs us about how much information is actually communicated across the channel when it is fed by the input distribution $P_{X}$. However, $I_{P_{X}}(X: Y)$ will not inform us about how much information we could in principle communicate across the channel, i.e., the capacity of the channel defined by:

$$
\begin{equation*}
\mathcal{C}=\underset{P_{X}}{\operatorname{argmax}} I_{P_{X}}(X: Y) . \tag{7}
\end{equation*}
$$

Thus, for describing the potential of a mechanism to modify information, we must define an information modification capacity in analogy to the definition of an information transmission capacity (Equation (7)) by maximizing the synergistic mutual information over all input distributions as:

$$
\begin{align*}
\mathcal{C}_{\bmod } & =\underset{P_{\mathbf{Y}^{-}, \mathbf{X}^{-}}}{\operatorname{argmax}} \tilde{I}_{\text {syn }}\left(Y^{+}: \mathbf{Y}^{-} ; \mathbf{X}^{-}\right),  \tag{8}\\
& =\underset{P_{\mathbf{Y}^{-}, \mathbf{x}^{-}}}{\operatorname{argmax}}\left[I\left(Y^{+}:\left(\mathbf{Y}^{-}, \mathbf{X}^{-}\right)\right)-\min _{Q \in \Delta_{P_{\mathbf{Y}^{+}, \mathbf{Y}^{-}, \mathbf{X}^{-}}}} I_{Q}\left(Y^{+}:\left(\mathbf{Y}^{-}, \mathbf{X}^{-}\right)\right)\right] . \tag{9}
\end{align*}
$$

How tractable the double optimization process implied in Equation (9) is in practice and whether analytical simplifications can be derived remains the topic of future work. However, other measures of PID that do not rely on an optimization over the space of probability distributions (such as the one by Finn et al. [40] in this special issue) may allow for the computation of a capacity for information modification-given the mechanism is known.

We would like to emphasize that maximizing synergy, or any other PID term, over possible input distributions is different from maximizing the same PID term via changes to the mechanism that yields the output, while keeping the input distributions fixed. This latter approach is considered in detail by the contribution of Rauh et al. [50] in this special issue.

### 4.5. On the Distinction between Information Modification and Noise

We emphasize that the definition of information modification used here (and first put forward in [11]) will not count information that is created de novo in an information processing element and then appears in its output. This is because modified information in the output has to be explained ultimately by the input to a processing element and the state (or history) of that element, taken together. This clearly does not hold for information just created independently of the processor's history. In other words, the information created de novo is counted as output noise instead of as modified information by our definition of information modification-a property that we consider desirable for any measure of information modification.

### 4.6. On the Relation between Transfer Entropy and Information Modification

As introduced in Section 2.2, the transfer entropy between two processes $\mathcal{X}, \mathcal{Y}$, where the variables $Y_{t}, X_{t}$ carry the current values of the processes and the variables $\mathbf{X}^{-}, \mathbf{Y}^{-}$carry the past state information is defined as [2]:

$$
\begin{equation*}
T E(\mathcal{X} \rightarrow \mathcal{Y})=I\left(Y_{t}: \mathbf{X}^{-} \mid \mathbf{Y}^{-}\right) \tag{10}
\end{equation*}
$$

As first noted by Williams and Beer [20], the (conditional) mutual information on the right-hand side can be decomposed using a PID as well. As shown in Section 2.2, this conditional mutual information is composed of both a unique contribution from the source, and a synergistic contribution where the current value $y_{t}$ is determined jointly-and not explainable in any simpler way-by the combination of past states $\mathbf{x}^{-}$and $\mathbf{y}^{-}$, i.e., the input from $\mathcal{X}$ to $\mathcal{Y}$ and $\mathcal{Y}^{\prime}$ s own history. (Of course, either component could be zero for a given distribution). Williams and Beer [10] suggested to call this synergistic part of the transfer entropy the (receiver) state dependent transfer entropy (SDTE) to highlight the interplay between sender and receiver in modifying the information. Obviously, such a subdivision of transfer is highly useful where computable. Naturally, the overlaps between the concepts of information modification and (multivariate) transfer entropy become more involved if $\mathcal{Y}$ receives more than one external input. What we label as information modification in this case would comprise the SDTE above, but perhaps not all of it, and also have additional contributions (see below and Figure 6).

This is a special case of the general effect that the synergistic components of a PID may change if additional inputs are considered, e.g., when the additional input on its own brings in information that is itself redundant with the information seen as synergistic between the other inputs. See, for example, the component labeled with $\left\{X_{2}\right\}\left\{M X_{1}\right\}$ in Figure 6, which is synergistic when not considering $X_{2}$, but redundantly also provided by $X_{2}$ alone. In more detail, a PID may decompose the information provided by a larger set of sources into many different shared (redundant), unique and synergistic components between subsets of these inputs. These components are placed onto a lattice (a partial order) by some variants of PID measures (see Section 4.1).


Figure 6. PID diagram for three input variables-two of them external inputs ( $X_{1}, X_{2}$ ), and one representing the past state of the receiving system $\left(M=\mathbf{Y}^{-}\right)$. The parts of the diagram highlighted in green would be considered information modification. These parts represent the information in the receiver that can only be explained by two or more input variables considered jointly.

### 4.7. New Research Perspectives in Neuroscience Based on PID and Information Modification

In closing, we would like to highlight the vast potential that PID and the analysis of information modification have both in understanding biological neural systems, and in designing artificial ones.

As detailed in [1], the comparison of shared vs. synergistic mutual information in the output of a neuron or neural network allows us to address directly issues of robust coding vs. maximizing coding capacity, and thereby helps us to understand fundamental design principles of biological networks.

Conversely, PID can also be used to define information theoretic goal functions and to derive local learning rules for neurons in artificial neural networks with unprecedented detail and precision as explicated in [17], extending popular information theoretic goal functions like infomax [51], or coherent infomax (see [52] and references therein). In particular, the formulation of novel PID estimators that no longer rely on an optimization step (see the work of Finn et al. [40] in this special issue) has seemingly removed remaining difficulties with an analytical treatment of this approach.

Moreover, the PID formalism lends itself easily to the analysis of both neural and behavioral data, enabling a direct comparison of the two. This will take our understanding of the relationship between neural activity and behavior beyond the level of an analysis of mere representations, i.e., beyond decoding representations of objects and intentions, to finding the loci of particular aspects of neural computation. For example, in a human performing a task requiring an XOR computation, one may look for hot-spots of synergistic mutual information in the system.

Ultimately, the ability to obtain a complete fingerprint of a neural computation in terms of active information storage, information transfer and, now, information modification makes it possible to identify algorithms implemented by a neural system-or at least strongly confines the search space. This finally allows to fully address the algorithmic level of understanding neural systems as formulated more than 30 years ago by David Marr ([53], also see [1]).

## 5. Conclusions

We used a recent extension of information theory here to measure where and when in a neural network information is not simply communicated through a channel but modified. The definition of information modification here builds on the concept of synergistic mutual information as introduced by Williams and Beer [10], and the measure defined by Bertschinger et al. [12]. We show that,
in the developing neural culture analyzed here, the contribution of synergistic mutual information rose as the network became more connected with development but ultimately dropped again as the activity became largely synchronized in bursts across the whole neural culture such that most mutual information was shared mutual information.

Acknowledgments: This project was supported through a Universities Australia/German Academic Exchange Service (DAAD) Australia-Germany Joint Research Cooperation Scheme grant (2016-17): "Measuring neural information synthesis and its impairment". Joseph T. Lizier was also supported through the Australian Research Council DECRA grant DE160100630, and a Faculty of Engineering and IT Early Career Researcher and Newly Appointed Staff Development Scheme 2016 grant. Viola Priesemann received financial support from the Max Planck Society, and from the German Ministry for Education and Research (BMBF) via the Bernstein Center for Computational Neuroscience (BCCN) Göttingen under Grant No. 01GQ1005B. The authors would like to thank all participants of the 2016 Workshop on Partial Information Decomposition held by the Frankfurt Institute for Advanced Studies (FIAS) and the Goethe University for their inspiring discussions on PID.
Author Contributions: Michael Wibral and Viola Priesemann conceived of the study; Michael Wibral and Viola Priesemann analyzed the data; Conor Finn, Joseph T. Lizier, Patricia Wollstadt and Michael Wibral wrote the code for PID computation in the IDT ${ }^{x l}$ toolbox; Michael Wibral, Viola Priesemann, and Joseph T. Lizier wrote the manuscript.
Conflicts of Interest: The authors declare no conflict of interest.

## References

1. Wibral, M.; Lizier, J.T.; Priesemann, V. Bits from Brains for Biologically Inspired Computing. Front. Robot. AI 2015, 2, 5.
2. Schreiber, T. Measuring information transfer. Phys. Rev. Lett. 2000, 85, 461-464.
3. Vicente, R.; Wibral, M.; Lindner, M.; Pipa, G. Transfer entropy-A model-free measure of effective connectivity for the neurosciences. J. Comput. Neurosci. 2011, 30, 45-67.
4. Wibral, M.; Pampu, N.; Priesemann, V.; Siebenhühner, F.; Seiwert, H.; Lindner, M.; Lizier, J.T.; Vicente, R. Measuring information-transfer delays. PLoS ONE 2013, 8, e55809.
5. Wollstadt, P.; Martínez-Zarzuela, M.; Vicente, R.; Díaz-Pernas, F.J.; Wibral, M. Efficient transfer entropy analysis of non-stationary neural time series. PLoS ONE 2014, 9, e102833.
6. Lizier, J.T.; Prokopenko, M.; Zomaya, A.Y. Local information transfer as a spatiotemporal filter for complex systems. Phys. Rev. E 2008, 77, 026110.
7. Lizier, J.T.; Prokopenko, M.; Zomaya, A.Y. Local measures of information storage in complex distributed computation. Inf. Sci. 2012, 208, 39-54.
8. Wibral, M.; Lizier, J.T.; Vögler, S.; Priesemann, V.; Galuske, R. Local active information storage as a tool to understand distributed neural information processing. Front. Neuroinf. 2014, 8, doi:10.3389/fninf.2014.00001.
9. Gomez, C.; Lizier, J.T.; Schaum, M.; Wollstadt, P.; Grützner, C.; Uhlhaas, P.; Freitag, C.M.; Schlitt, S.; Bölte, S.; Hornero, R.; et al. Reduced Predictable Information in Brain Signals in Autism Spectrum Disorder. Front. Neuroinf. 2014, 8, 9.
10. Williams, P.L.; Beer, R.D. Nonnegative Decomposition of Multivariate Information. arXiv 2010, arXiv:1004.2515.
11. 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.
12. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying Unique Information. Entropy 2014, 16, 2161-2183.
13. Griffith, V.; Koch, C. Quantifying Synergistic Mutual Information. In Guided Self-Organization: Inception; Prokopenko, M., Ed.; Springer: Berlin, Germany, 2014; Volume 9, pp. 159-190.
14. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. Phys. Rev. E 2013, 87, 012130.
15. Fano, R. Transmission of Information; The MIT Press: Cambridge, MA, USA, 1961.
16. Cover, T.M.; Thomas, J.A. Elements of Information Theory; Wiley-Interscience: New York, NY, USA, 1991.
17. 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 Cognit. 2017, 112, 25-38.
18. Langton, C.G. Computation at the edge of chaos: Phase transitions and emergent computation. Phys. D Nonlinear Phenom. 1990, 42, 12-37.
19. Mitchell, M. Computation in Cellular Automata: A Selected Review. In Non-Standard Computation; Gramß, T., Bornholdt, S., Groß, M., Mitchell, M., Pellizzari, T., Eds.; Wiley-VCH Verlag GmbH \& Co. KGaA: Weinheim, Germany, 1998; pp. 95-140.
20. Williams, P.L.; Beer, R.D. Generalized Measures of Information Transfer. arXiv 2011, arXiv:1102.1507.
21. Timme, N.M.; Ito, S.; Myroshnychenko, M.; Nigam, S.; Shimono, M.; Yeh, F.C.; Hottowy, P.; Litke, A.M.; Beggs, J.M. High-Degree Neurons Feed Cortical Computations. PLoS Comput. Biol. 2016, 12, 1-31.
22. Wollstadt, P.; Lizier, J.T.; Finn, C.; Martinz-Zarzuela, M.; Vicente, R.; Lindner, M.; Martinez-Mediano, P.; Wibral, M. The Information Dynamics Toolkit, IDT ${ }^{x l}$. Available online: https://github.com/pwollstadt/IDTxl (accessed on 25 August 2017).
23. Makkeh, A.; Theis, D.O.; Vicente Zafra, R. Bivariate Partial Information Decomposition: The Optimization Perspective. 2017, under review.
24. Wagenaar, D.A.; Pine, J.; Potter, S.M. An extremely rich repertoire of bursting patterns during the development of cortical cultures. BMC Neurosci. 2006, 7, 11.
25. Wagenaar, D.A. Network Activity of Developing Cortical Cultures In Vitro. Available online: http: / /neurodatasharing.bme.gatech.edu/development-data/html/index.html (accessed on 25 August 2017).
26. Wagenaar, D.; DeMarse, T.B.; Potter, S.M. MeaBench: A toolset for multi-electrode data acquisition and on-line analysis. In Proceedings of the 2nd International IEEE EMBS Conference on Neural Engineering, Arlington, VA, USA, 16-20 March 2005; pp. 518-521.
27. Timme, N.; Shinya, I.; Maxym, M.; Fang-Chin, Y.; Emma, H.; Pawel, H.; Beggs, J.M. Multiplex Networks of Cortical and Hippocampal Neurons Revealed at Different Timescales. PLoS ONE 2014, 9, 1-43.
28. Faes, L.; Marinazzo, D.; Montalto, A.; Nollo, G. Lag-specific transfer entropy as a tool to assess cardiovascular and cardiorespiratory information transfer. IEEE Trans. Biomed. Eng. 2014, 61, 2556-2568.
29. Lizier, J.T.; Rubinov, M. Inferring effective computational connectivity using incrementally conditioned multivariate transfer entropy. BMC Neurosci. 2013, 14, P337.
30. Montalto, A.; Faes, L.; Marinazzo, D. MuTE: A MATLAB Toolbox to Compare Established and Novel Estimators of the Multivariate Transfer Entropy. PLoS ONE 2014, 9, e109462.
31. Lindner, M.; Vicente, R.; Priesemann, V.; Wibral, M. TRENTOOL: A Matlab open source toolbox to analyse information flow in time series data with transfer entropy. BMC Neurosci. 2011, 12, 1-22.
32. Wollstadt, P.; Meyer, U.; Wibral, M. A Graph Algorithmic Approach to Separate Direct from Indirect Neural Interactions. PLoS ONE 2015, 10, e0140530.
33. Levina A.; Priesemann V. Subsampling scaling. Nat. Commun. 2017, 8, 15140.
34. Levina, A.; Priesemann, V. Subsampling Scaling: A Theory about Inference from Partly Observed Systems. arXiv 2017, arXiv:1701.04277.
35. Priesemann, V.; Wibral, M.; Valderrama, M.; Pröpper, R.; le van Quyen, M.; Geisel, T.; Triesch, J.; Nikolić, D.; Munk, M.H.J. Spike avalanches in vivo suggest a driven, slightly subcritical brain state. Front. Syst. Neurosci. 2014, 8, 108.
36. Priesemann, V.; Lizier, J.; Wibral, M.; Bullmore, E.; Paulsen, O.; Charlesworth, P.; Schröter, M. Self-organization of information processing in developing neuronal networks. BMC Neurosci. 2015, 16, P221.
37. Timme, N.; Alford, W.; Flecker, B.; Beggs, J.M. Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. J. Comput. Neurosci. 2014, 36, 119-140.
38. Phillips, W.A. Cognitive functions of intracellular mechanisms for contextual amplification. Brain Cognit. 2017, 112, 39-53.
39. Larkum, M. A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. Trends Neurosci. 2013, 36, 141-151.
40. Finn, C.; Prokopenko, M.; Lizier, J.T. Pointwise Partial Information Decomposition Using the Specificity and Ambiguity Lattices. 2017, under review.
41. 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: Cham, Vietnam, 2013; pp. 251-269.
42. Ince, R.A.A. Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. Entropy 2017, 19, 318.
43. 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, Honolulu, HI, USA, 29 June-4 July 2014; pp. 2232-2236.
44. Pica, G.; Piasini, E.; Chicharro, D.; Panzeri, S. Invariant Components of Synergy, Redundancy, and Unique Information among Three Variables. Entropy 2017, 19, 451.
45. Olbrich, E.; Bertschinger, N.; Rauh, J. Information Decomposition and Synergy. Entropy 2015, 17, 3501-3517.
46. Schneidman, E.; Puchalla, J.L.; Segev, R.; Harris, R.A.; Bialek, W.; Berry, M.J. Synergy from Silence in a Combinatorial Neural Code. J. Neurosci. 2011, 31, 15732-15741.
47. Stramaglia, S.; Wu, G.R.; Pellicoro, M.; Marinazzo, D. Expanding the transfer entropy to identify information circuits in complex systems. Phys. Rev. E 2012, 86, 066211.
48. Stramaglia, S.; Cortes, J.M.; Marinazzo, D. Synergy and redundancy in the Granger causal analysis of dynamical networks. New J. Phys. 2014, 16, 105003.
49. Stramaglia, S.; Angelini, L.; Wu, G.; Cortes, J.M.; Faes, L.; Marinazzo, D. Synergetic and redundant information flow detected by unnormalized Granger causality: Application to resting state fMRI. IEEE Trans. Biomed. Eng. 2016, 63, 2518-2524.
50. Rauh, J.; Banerjee, P.; Olbrich, E.; Jost, J.; Bertschinger, N. On Extractable Shared Information. Entropy 2017, 19, 328.
51. Linsker, R. Self-organisation in a perceptual network. IEEE Comput. 1988, 21, 105-117.
52. Kay, J.W.; Phillips, W.A. Coherent Infomax as a computational goal for neural systems. Bull. Math. Biol. 2011, 73, 344-372.
53. Marr, D. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information; Henry Holt and Co., Inc.: New York, NY, USA, 1982.

[^0]:    Wollstadt et al., (2019). IDTxI: The Information Dynamics Toolkit xl: a Python package for the efficient analysis of multivariate information 1 dynamics in networks. Journal of Open Source Software, 4(34), 1081. https://doi.org/10.21105/joss. 01081

[^1]:    Wollstadt et al., (2019). IDTxI: The Information Dynamics Toolkit xl: a Python package for the efficient analysis of multivariate information 2 dynamics in networks. Journal of Open Source Software, 4(34), 1081. https://doi.org/10.21105/joss. 01081

[^2]:    Wollstadt et al., (2019). IDTxI: The Information Dynamics Toolkit xI: a Python package for the efficient analysis of multivariate information 3 dynamics in networks. Journal of Open Source Software, 4(34), 1081. https://doi.org/10.21105/joss.01081

