GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Information, Logic, and Inference in the Analysis of Complex Networks

Dissertation
for the award of the degree
“Doctor of Philosophy”
Division of Mathematics and Natural Sciences
of the Georg-August-Universität Göttingen

within the doctoral program *Theoretical and Computational Neuroscience*
of the Georg-August University School of Science (GAUSS)

submitted by
Aaron Julian Gutknecht

from Lich
Göttingen 2023

Thesis Advisory Committee:

Prof. Dr. Michael Wibral, Campus Institute for Dynamics of Biological Networks, Georg-August University Göttingen

Prof. Dr. Fred Wolf, Campus Institute for Dynamics of Biological Networks, Georg-August University Göttingen & Department of Nonlinear Dynamics, Max Planck Institute for Dynamics and Self-Organization

Dr. Lionel Barnett, Department of Informatics, University of Sussex, Brighton, UK

Members of the Examination Board:

Referee: Prof. Dr. Michael Wibral, Campus Institute for Dynamics of Biological Networks, Georg-August University Göttingen

2nd Referee: Prof. Dr. Fred Wolf, Campus Institute for Dynamics of Biological Networks, Georg-August University Göttingen & Department of Nonlinear Dynamics, Max Planck Institute for Dynamics and Self-Organization

Further members of the Examination Board:

Dr. Lionel Barnett, Department of Informatics, University of Sussex, Brighton, UK

Prof. Dr. Alexander Ecker, Department of Computer Science, Georg-August University Göttingen

Prof. Dr. Ulrich Parlitz, Department of Biomedical Physics, Max Planck Institute for Dynamics and Self-Organization

Prof. Dr. Florentin Wörgötter, Third Institute of Physics - Biophysics, Georg-August University Göttingen

Date of oral examination: 04.12.2023

Acknowledgement

First and foremost, I wish to express my heartfelt gratitude to my primary advisor, Prof. Michael Wibral. His guidance and unwavering support have been instrumental in shaping my academic journey. I particularly enjoyed the academic freedom he provided, allowing me to explore diverse disciplines, poke at unusual ideas, and go down the occasional rabbit hole.

I am also deeply thankful to my co-advisor Dr. Lionel Barnett for his guidance and the memorable experiences during my visits to the University of Sussex. Our intellectually stimulating discussions have played a significant role in my academic growth.

I also wish to extend special thanks to my co-advisor Prof. Fred Wolf for his valuable insights and supervision during the Thesis Advisory Committee (TAC) meetings.

I am very fortunate to have had Dr. Abdullah Makkeh as both a collaborator and coworker. His consistent support and encouragement have been exemplary.

I'm also indebted to Prof. Anil Seth and Prof. Jürgen Jost for hosting me at the University of Sussex and the Max Planck Institute for Mathematics in the Sciences, respectively.

Moreover, I would like to acknowledge Prof. Thomas Metzinger for sparking my fascination with research across disciplinary boundaries and Prof. André Fuhrmann for igniting my interest in formal logic, both of which laid the foundation for my PhD research.

I'm grateful to my colleagues both at the MEG Lab in Frankfurt and the Wibral Lab in Göttingen—namely, Anya Dietrich, Edoardo Pinzutti, Cora Fisher, Andreas Schneider, David Ehrlich, and Kyle Schick-Poland—for creating supportive and enriching work environments.

On a more personal note, I owe immense gratitude to my family—my mother Dorothee, my father Wolfram, my brother Milan as well as my partner Alina—for their love, encouragement, and support throughout this journey. A special mention is due to my grandmother Elke, whose support has been invaluable in advancing my academic pursuits. Lastly, I am grateful to David and Patrick for their friendship and the many shared experiences that brought joy and a sense of balance during my PhD studies, and to Marjorie, my friend and English teacher, whose lessons on academic writing became a cornerstone for my path in academia.

Summary

The following thesis deals with a range of current topics in information theory and statistics. It consists of five distinct contributions: Chapter 2 focuses on the statistics of single-regression Granger causality estimators. Chapters 3-5 address the theory of Partial Information Decomposition (PID), an extension of classical Shannon Information Theory. Chapter 6 is about Significant Subgraph Mining, a statistical method for finding differences between graph-generating processes with multiple comparisons correction. In the following, a brief summary of each contribution is provided:

Chapter 2, "Sampling Distribution of Single Regression Granger Causality estimators", deals with the statistics of single regression Granger causality estimators for which only the full auto-regressive model has to be estimated while the parameters of the reduced model (regressing the target process only on its own past) are analytically or numerically derived from the full model parameters. This is in contrast to standard dual regression estimators for which both the full and the reduced model have to be estimated. The paper shows that the asymptotic distribution of single regression Granger causality estimators under the hypothesis of vanishing Granger causality is a generalized χ^2-distribution which is in many cases well approximated by a Γ-distribution. This is true for time-domain Granger causality as well as band-limited Granger causality which is particularly useful for neuroscientific applications in which a particular frequency-band may be of interest. The paper also derives asymptotically valid significance tests based on the derived sampling distributions.

Chapter 3, "Introducing a differentiable measure of pointwise shared information", proposes a measure of the information shared between particular realizations of a set of source variables about a particular realization of a target variable. In this sense it is a pointwise measure. It is constructed in close analogy to classical pointwise mutual information. This can be achieved in two ways: First, based on the insight that pointwise mutual information can be defined in terms of probability mass exclusions. Analogously, pointwise shared information may be introduced in terms of shared probability mass exclusions. Second, pointwise mutual information can be seen as the information about the value of a target variable provided by the truth of a certain logical statement about the source variables. Similarly, there is a logical statement about the source realizations that reasonably carries their shared information about

the target realization. The resulting measure of pointwise shared information i^{sx} exhibits desirable properties for applications, in particular its differentiability with respect to the underlying probability distribution. Further, any general measure of shared information implies an entire Partial Information Decomposition, which in the case of i^{sx} will also be differentiable. This makes it possible to define goal functions in terms of PID quantities (e.g. "maximize redundancy") with which neural networks can be trained.

Chapter 4, "Bits and Pieces: understanding information decomposition from part-whole relations and formal logic", shows that the entire theory of PID can be derived, firstly, from considerations of part-whole relationships between information atoms and mutual information terms, and secondly, based on a hierarchy of logical constraints describing how a given information atom can be accessed. In this way, the idea of a PID is developed on the basis of two of the most elementary relationships in nature: the part-whole relationship and the relation of logical implication. This unifying perspective provides insights into pressing questions in the field such as the possibility of constructing a PID based on concepts other than redundant information in the general n-sources case. The paper also presents a re-derivation of the shared exclusions measure of redundant information introduced in Chapter 3 based on principles of logic and mereology (the study of part-whole relationships).

Chapter 5, "From Babel to Boole: The Logical Structure of Information Decompositions", expands upon the ideas presented in "Bits and Pieces". The central theme of this chapter revolves around PID "base-concepts". These are information functionals which, when defined, induce a complete PID. Within the parthood approach, these base-concepts are expressed in terms of conditions phrased in formal logic on the specific parthood relations between the PID components and the different mutual information terms. The work identifies a general pattern for these logical conditions. Every PID base-concept in the existing literature fits within this pattern as special cases. Moreover, it leads to a novel base-concepts called "vulnerable information" which quantifies information that may be lost if one loses access to one of the sources. Furthermore, all PID base-concepts are shown to fall into equivalence classes of measures that describe the same information components but viewed from the perspective of different source collections.

Chapter 6, "Significant Subgraph Mining for Neural Network Analysis with Multiple Comparison Correction", addresses a problem of graph statistics which often comes up in the next step after an information theoretic analysis. Suppose for instance that we have performed a pairwise Granger causality analysis of MEG data in two experimental groups. For each group we obtain a set of graphs (one for each sub-

ject) and we would like to know if there are any differences between the groups. Maybe a particular connection is more likely to occur in one group rather than the other. And even if there are no such differences on a per-link basis, there may be differences in the dependencies between links. For instance, while two connections may always appear together in one group they may occur completely independently in the other. In principle, any possible stochastic difference between the two graph graph-generating processes can be expressed in terms of the probabilities of occurrence of specific subgraphs. Significant Subgraph Mining systematically tests all such differences while correcting for the formidable multiple comparisons problem arising because the total number of possible subgraphs scales super-exponentially in the number of graph nodes. The paper extends the method to within-subject experimental designs that allows for dependencies between the graph-generating processes. It also provides a systematic analysis of its error-statistical properties in simulation and in empirical data in order to derive practical recommendations for the application of subgraph mining in neuroscience. In particular, it presents an empirical power analysis for Transfer Entropy networks inferred from resting state MEG data comparing autism spectrum patients with neurotypical controls. Finally, a python implementation as part of the openly available IDTxl toolbox is provided.

Contents

1 Introduction

Since its inception in the mid-20th century, information theory has evolved into an indispensable framework for deciphering the principles of information flow and processing in a diverse range of scientific fields. Conceived originally by Claude Shannon for telecommunications, the theory's foundational concepts have permeated disciplines as diverse as biology, computer science, and economics.

This thesis is concerned with the theoretical and statistical underpinnings of information theory. It explores three major topics:

1. The statistical theory of linear information flow, framed in terms of Granger-Geweke causality (Chapter 2). In this part, the thesis analytically derives the sampling distribution for an efficient class of Granger causality estimators, known as 'single regression estimators,' in both time and frequency domains. The work also constructs statistical tests based on these derived distributions.

2. The theory of Partial Information Decomposition (PID) as a tool to overcome apparent limitations of classical information theory in providing a comprehensive picture of the informational relationships between multiple variables (Chapters 3-5). In this part, the thesis employs insights from formal logic and mereology (the study of part-whole relationships) to offer a unifying perspective on the mathematical structure underlying the PID problem, and presents a concrete solution.

3. The extension, adaptation, and software implementation of Significant Subgraph Mining for comparative analysis of neural networks inferred via information theoretic measures of functional connectivity (Chapter 6).

Overall, this thesis aims to advance our theoretical understanding and statistical inference methods pertaining to the flow and processing of information in diverse complex systems. It does so by synthesizing insights from multiple disciplines to tackle foundational questions in a novel way. By addressing both theoretical principles and practical statistical methods, the thesis fills key gaps in existing literature, providing a more comprehensive toolkit for researchers studying the complex interdependencies governing biological as well as non-biological networks.

In the following sections, we explore the necessary theoretical background, the scientific relevance, and the specific contributions of the work presented in subsequent chapters. We begin with a brief overview of classical Shannon information theory as it applies to 'static' variables without consideration of time. From there, we transition to the realm of information dynamics, describing the informational relations between stochastic processes. Key measures such as Transfer Entropy and its linear approximation, Granger-Geweke causality, are introduced. Subsequently, we review the foundational concept of Partial Information Decomposition, an extension of classical information theory, and how it can be applied to a variety of questions in complex systems, neuroscience and beyond. Finally, we explore how information theoretic analyses often lead to graph structures describing patterns of information flow. This creates a need for statistical techniques that can effectively compare these graph structures between groups or experimental conditions while also handling the severe multiple comparisons problem arising in this context.

1.1 A brief tour of classical information theory

In his landmark paper, 'A Mathematical Theory of Communication,' Claude Shannon formulated three axioms that any reasonable measure of the information in a random variable X should satisfy [1]. Viewing information as a measure of uncertainty about the value of X, he proposed that

1. The information of X should be a continuous function of the underlying probability distribution. Small changes in $P_X(x)$ should not drastically change our uncertainty about X.

2. If the probability distribution is uniform, i.e., $P_X(x) = \frac{1}{m}$, then the information content in X should be monotonically increasing with the size of the alphabet m. The more equally likely values X can take, the more uncertain we are about its value.

3. If the choice between the different values of X is broken down into multiple choices, the information of X can be computed as a weighted average of individual information contributions associated with those choices. Specifically, if we partition the values of X into multiple groups, our uncertainty about X can be expressed as our uncertainty about which group X falls into plus our uncertainties about which value X takes given that it falls within a particular group (weighted by the probabilities of the respective groups).

These axioms lead to the following unique expression for measuring the information in X (up to some arbitrary constants):

$$H(X) = -\sum_x P_X(x) \log P_X(x) \tag{1.1}$$

This quantity is known as the *entropy* of X, and it serves as the foundational building block for more complex informational quantities (for a good textbook on information theory see [2]).

Having established the concept of entropy for a single variable, it is natural to extend our understanding to scenarios involving multiple variables. Specifically, consider a situation where we observe the value y of another variable Y. How does this observation affect our uncertainty about X? This leads us to the notion of *conditional entropy of X given $Y = y$*:

$$H(X|Y = y) = -\sum_x P_{X|Y}(x|y) \log P_{X|Y}(x|y) \tag{1.2}$$

Formally, this is the same as the unconditional entropy but with all distributions conditioned on $Y = y$. Intuitively, it is the uncertainty about X in the conditional universe where we know that Y has assumed the value y. If X and Y are independent this reduces to the unconditional entropy. Our uncertainty about X is just as great as it was before having observed Y in this case. Averaging this quantity over all values of Y results in the *conditional entropy of X given Y*:

$$H(X|Y) = -\sum_{x,y} P_{XY}(x, y) \log P_{X|Y}(x|y) \tag{1.3}$$

Given these definitions and their interpretations it is straightforward to introduce a measure of the information that X provides about Y, the *mutual information*:

$$I(X : Y) = H(Y) - H(Y|X) \tag{1.4}$$

It compares our uncertainty about Y before having observed X with our average uncertainty after having observed X. In other words, it is our average *reduction* of uncertainty about Y after observing the value of X. In many contexts it is crucial to also consider the *conditional mutual information* that X provides about Y given a third variable Z:

$$I(X : Y|Z) = H(Y|Z) - H(Y|X, Z) \tag{1.5}$$

So essentially we are putting ourselves in a universe where the value of the third variable Z is known and ask, given this knowledge, how much further is our uncertainty about Y reduced by additionally observing X.

Mutual information has many insightful properties of which we will here only mention the ones which are most important in the remainder of this thesis. *Firstly*, mutual information is *symmetric*:

$$I(X:Y) = I(Y:X) \tag{1.6}$$

The information provided by X about Y is the same as the information provided by Y about X. *Secondly*, it follows from Jensen's inequality that mutual information is non-negative

$$I(X:Y) \geqslant 0 \tag{1.7}$$

There can be no misinformation between random variables if information is understood in the sense of mutual information. *Thirdly*, mutual information is equal to zero if and only if X and Y are independent

$$I(X:Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y \tag{1.8}$$

In fact mutual information can be equivalently defined as a measure of divergence of the joint distribution $P_{X,Y}$ from the product distribution $P_X P_Y$, representing the case of independence. This may also provide some intuition on the non-negativity of mutual information: it only measures how strongly dependent two variables are, irrespective of the type of the dependence. Similarly, *conditional* mutual information measures how strongly dependent two variables are given a third variable. Accordingly, it is zero just in case X and Y are conditionally independent given Z:

$$I(X:Y|Z) = 0 \Leftrightarrow X \perp\!\!\!\perp Y|Z \tag{1.9}$$

It is important to point out that conditional mutual information may be larger, smaller or equal to conditional mutual information. This is rooted in the underlying fact from probability theory that independence and conditional independence are logically distinct concepts in the sense that neither implies the other.

Forthly, recall that X and Y could be vectors with multiple components. Mutual information satisfies a chain rule expressing how the mutual information of multiple

variables $X_1, \ldots, X_n$ about another variable Y can be broken down into simpler mutual information terms:

$$I(X_1, \ldots, X_n : Y) = \sum_{i=1}^{n} I(X_i : Y | X_1, \ldots, X_{i-1}) \tag{1.10}$$

Intuitively: the information provided by all the X_i about Y is the information about Y provided by X_1 plus the additional information provided by X_2 about Y given that we already know X_1 plus the additional information provided by X_3 about Y given that we already know X_1 and X_2 and so forth.

Fifthly, mutual information satisfies the so called Data Processing Inequality. Suppose we start with variables X and Y that might be statistically dependent. $I(X : Y)$ quantifies this dependence. Now, suppose further that Y is processed in some way to obtain a third variable Z. This processing may be deterministic, i.e. $Z = f(Y)$ for some deterministic function f, or it may be probabilistic. For instance, we might have $Z = f(Y, N)$ where f is again some function and N is some independent noise variable. The Data Processing Inequality states that no such processing can increase the amount of information X has about Y. In full generality the statement says that if $X - Y - Z$ form a Markov-Chain, i.e. Z is conditionally independent of X given Y, then

$$I(X : Y) \geqslant I(X : Z) \tag{1.11}$$

Finally, a last property of mutual information that deserved some attention is its invariance under invertible tranformations of the variables. Specifically, let f and g be bijective functions on the codomains of variables X and Y. Then, we have

$$I(f(X) : g(Y)) = I(X : Y) \tag{1.12}$$

Intuitively, the information provided by one variable about another should not depend on the way we describe it, e.g. the units in which we measure it. Converting everything from, say, meters to kilometers, or even from meters to $\log$-meters, should not affect how much information X carries about Y. That this is indeed the case can be shown immediately based on the Data Processing Inequality using f and g as the processing functions.

The transformation invariance of mutual information only becomes truly interesting in the continuous case. For discrete variables the transformations amount to a simple relabelling of the values of the variables (e.g. "A","B","C" instead of "1","2","3"). This however will not affect the expressions for the entropies (Eqs. 1.1-1.2) at all. In the continuous case, the situation is more complicated because transformations can greatly affect the shape of the joint distributions of the variables. Hence, it is much

more surprising, and hence more profound, that mutual information would still remain invariant in the continuous case.

Mutual information can be defined for jointly distributed continuous variables in much the same way as discrete mutual information by simply replacing sums by integrals and probability mass functions by probability densities in the entropy expressions 1.1-1.2 leading to *differential entropies*:

$$h(X) = -\int f(x)\log(f(x))dx \qquad h(X|Y) = -\int f(x,y)\log(f(x|y))dxdy \tag{1.13}$$

The differential entropies themselves behave quite differently from discrete entropies. In particular, they can be negative and are thus not easily interpretable as measures of uncertainty about a variable (or uncertainty about a variable given observations of another variable). However, the continuous mutual information

$$I(X:Y) = h(X) - h(X|Y) \tag{1.14}$$

retains all of the properties we discussed above.

A last information theoretic quantity that deserves a short introduction since it plays a role in Chapters 3-5 about Partial Information Decomposition is the *interaction information*. Recall that we introduced the mutual information as the difference between an unconditional entropy and a conditional entropy. One can continue this pattern to recursively construct "higher-order" information quantities by subtracting unconditional and unconditional quantities of lower order. The next step in the sequence would be the difference between unconditional mutual information and conditional mutual information. This is generally known as the interaction information [3]

$$I(X:Y:Z) = I(X:Y) - I(X:Y|Z) \tag{1.15}$$

Similar to how mutual information is expressed as a change in entropy upon observing a second variable, interaction information is the change in mutual information upon observing a third one. However, interaction information may be negative because the existence of dependence between variables doesn't automatically imply the presence or absence of conditional independence when a third variable is introduced, and vice versa.

1.2 Information dynamics

1.2.1 Basic concepts

So far we have only discussed the informational relationships between "static" variables without an explicit time dimension. However, in the analysis of complex systems it is often their time evolution which is of particular interest and data come in the form of time series. Particularly in neuroscience such data are pervasive, for instance in the form of magnetoencephalography (MEG), electroencephalography (EEG) or functional magnetic resonance imaging (fMRI) data. Such data are best described as being generated by *stochastic processes* $\mathbf{X}_t$, i.e. infinite sequences of random vectors indexed with a time index $t \in \mathbb{Z}$ (the time index may be continuous but we are focussing on the discrete time case here).

The study of informational relationships between stochastic processes has been termed *information dynamics* [4]. There are in particular two well established information quantities of interest in this context: *Active Information Storage* and *Transfer Entropy* [5, 6]. Both of these quantities seek to measure the degree to which the future of the process can be predicted. The AIS quantifies the predictability of $\mathbf{X}_{t+1}$ based on the entire history of the process itself:

$$AIS_{\mathbf{X}}(t) = I(\mathbf{X}_{t+1} : \mathbf{X}_t^-) \tag{1.16}$$

where $\mathbf{X}_t^- = \mathbf{X}_t, \mathbf{X}_{t-1}, \ldots$ is the history of the process up to and including the present. In this way it describes the information stored in the process about what it will do next. The transfer entropy quantifies the predictability of the process $\mathbf{X}_t$ based on another process $\mathbf{Y}_t$ *over and above* the ability of $\mathbf{X}_t$ to self-predict. It is thus introduced as a conditional mutual information:

$$TE_{\mathbf{Y}\to\mathbf{X}}(t) = I(\mathbf{X}_{t+1} : \mathbf{Y}_t^- | \mathbf{X}_t^-) \tag{1.17}$$

A conditional version of this, taking into account other potential predictors $\mathbf{Z}_t$, can easily be introduced and is often used in practice:

$$TE_{\mathbf{Y}\to\mathbf{X}|\mathbf{Z}}(t) = I(\mathbf{X}_{t+1} : \mathbf{Y}_t^- | \mathbf{X}_t^-, \mathbf{Z}_t^-) \tag{1.18}$$

The time index in the definition of AIS and TE is necessary if we are considering general stochastic processes. However, if we are restricting ourselves to *strongly stationary* processes it can be removed. Such processes have the property that the joint distribution of any finite subselection of process variables is invariant

under time shifts. In other words if we are considering variables $\mathbf{X}_{t_1}, \mathbf{X}_{t_2}, \ldots, \mathbf{X}_{t_n}$, their joint distribution does not dependent on the specific time points $t_1, \ldots, t_n$ but only their temperal relations. For example, in a strongly stationary process X_5, X_7, X_{13} would have the same joint distributions as X_1, X_3, X_9 since these two sets of variables are simply shifted by a time lag of 4.

Transfer Entropy stands in an intimate relation to a concept of causality between processes introduced by Clive Granger and influenced by earlier works by Norbert Wiener [7]. According to this conception a process $\mathbf{Y}_t$ "causes" a process $\mathbf{X}_t$ just in case $\mathbf{X}_t$ becomes more predictable when taking into account the history of $\mathbf{Y}_t$ in addition to the history of $\mathbf{X}_t$ itself and the history of other potentially explanatory variables $\mathbf{Z}_t$. Granger argues that the more carefully $\mathbf{Z}_t$ is chosen so that it alone provides as much predictability of $\mathbf{X}_t$ as possible, the more stringently the "causality" from $\mathbf{Y}_t$ to $\mathbf{X}_t$ is tested. If $\mathbf{Y}_t$ still provides some unique additional insight, then, according to Granger, the term "causality" is justified [8].

Granger's condition of causality is often expressed in terms of a conditional dependence statement [5]:

$$\mathbf{X}_{t+1} \not\perp\!\!\!\perp \mathbf{Y}_t^- \mid \mathbf{X}_t^-, \mathbf{Z}_t^- \tag{1.19}$$

This states that the future of $\mathbf{X}$ is conditionally dependent of the history of $\mathbf{X}$ given its own history and the history of other explanatory variables $\mathbf{Z}$. The transfer entropy being a conditional mutual information this is equivalent to (due to property 1.9 above):

$$TE_{\mathbf{Y}\to\mathbf{X}|\mathbf{Z}} \neq 0 \tag{1.20}$$

The term Granger causality is not only used to refer to the *condition* for causality just described but also to a *measure* of this causal influence for linear stochastic processes, i.e. a function that assigns a numerical value to the Granger-causal influence from $\mathbf{Y}_t$ to $\mathbf{X}_t$. The most influential formulation of such a measure is that of Geweke [9–11] in the context of vector-autoregressive (VAR) models [12]. It expresses Granger causality in terms of the reduction of the residual's variance when comparing a reduced linear regression of $\mathbf{X}_t$ on its own past with a full linear regression that additionally regresses on the past of $\mathbf{Y}_t$ (full details are given in Chapter 2).

More formally, consider a vector stochastic process $(\mathbf{X}_t, \mathbf{Y}_t)$ where $\mathbf{X}_t$ has two stable, possibly infinite-order, VAR representations: The full model

$$\mathbf{X}_t = \sum_{p=1}^{\infty} A_{xx,p}\mathbf{X}_{t-p} + A_{xy,p}\mathbf{Y}_{t-p} + \boldsymbol{\epsilon}_{x,t} \tag{1.21}$$

and the reduced model

$$\mathbf{X}_t = \sum_{p=1}^{\infty} A^R_{xx,p} \mathbf{X}_{t-p} + \boldsymbol{\epsilon}^R_{x,t} \tag{1.22}$$

such that the residuals $\boldsymbol{\epsilon}_{x,t}$ and $\boldsymbol{\epsilon}^R_{x,t}$ in both VAR representations are white noise processes (zero mean and temporarily uncorrelated) with positive definite covariance matrices Σ and Σ^R. Under these conditions, the Granger-Geweke causality from $\mathbf{Y}$ to $\mathbf{X}$ is defined as:

$$F_{\mathbf{X}\to\mathbf{Y}} = \log \frac{|\Sigma^R|}{|\Sigma|} \tag{1.23}$$

where $|\circ|$ denotes the determinant. From an information theoretic perspective this linear measure of improved predictability can be thought of as a linear approximation to the more general non-linear Transfer Entropy. This is rooted in the fact that for Gaussian processes (where dependencies are completely described by correlations) the two measures are equivalent [13].

One significant advantage of Granger causality over Transfer Entropy lies in its relative ease of estimation and testing using empirical data. While Transfer Entropy and other information-theoretic measures are notoriously difficult to estimate—often requiring large datasets for accurate results—Granger causality allows for a more straightforward analysis. Specifically, standard statistical methods, such as maximum likelihood estimation and likelihood ratio testing, can be readily applied in the case of parametric Granger causality. However, it is important to note that this comes at the cost of being constrained to a more restrictive underlying model class.

Despite these relative advantages of Granger causality, it is crucial to recognize that its estimation and testing come with their own set of challenges. Specifically, dual regression estimators, which require separate estimation of the full and reduced autoregressive models and then plug the estimated residual covariances into the Granger causality formula, suffer from a severe bias-variance trade-off (see in particular the instructive exchange described in [14–17]). In response to these issues, more efficient alternatives known as 'single regression estimators' have been developed [18]. Unlike their dual regression counterparts, these estimators require only a single estimation of the full autoregressive model. The required reduced model residuals' covariance is then calculated analytically (or numerically to desired precision) based on the estimated full model parameters.

1.2.2 Contribution of this work

Although single regression estimators offer a more efficient alternative to dual regression methods in Granger causality analysis, an important open question has persisted in the literature: the absence of an established asymptotic theory for these estimators. Chapter 2 of this thesis ("Sampling distribution for Single Regression Granger Causality Estimators") closes this gap by deriving the asymptotic sampling distributions for single regression Granger causality estimators in both the time and frequency domains. Subsequently, the chapter introduces valid statistical tests based on these newly derived distributions. Finally, the chapter outlines how the same methods can also be utilized to derive the sampling distribution under the alternative hypothesis, for the conditional case and for general state-space models.

1.3 Partial Information Decomposition

1.3.1 The PID problem

Partial Information Decomposition is an extension of classical information theory that promises a more fine grained picture of the informational relationships between variables. It was originally proposed in a seminal paper by Williams and Beer in 2010 [19]. In the most basic case it considers two random variables S_1 and S_2 called the *information sources* and a random variable T called the *target*. All variables are jointly distributed and the goal is to decompose the mutual information carried by the sources about the target $I(S_1, S_2 : T)$ into four components:

1. The information $\Pi(\{1\} : T)$ provided *uniquely* by S_1 about T.

2. The information $\Pi(\{2\} : T)$ provided *uniquely* by S_2 about T.

3. The information $\Pi(\{1\}\{2\} : T)$ provided *redundantly* by both S_1 and S_2.

4. The information $\Pi(\{1, 2\} : T)$ provided *synergistically* by S_1 and S_2 which is only revealed when both variables are known at the same time.

Intuitively, these components, often called the *information atoms*, should stand in the following relations to mutual information terms

$$I(S_1, S_2 : T) = \Pi(\{1\}\{2\} : T) + \Pi(\{1\} : T) + \Pi(\{2\} : T) + \Pi(\{1, 2\} : T) \tag{1.24}$$

$$I(S_1 : T) = \Pi(\{1\}\{2\} : T) + \Pi(\{1\} : T) \tag{1.25}$$

$$I(S_2 : T) = \Pi(\{1\}\{2\} : T) + \Pi(\{2\} : T) \tag{1.26}$$

This is illustrated visually in Figure 1.1. Unfortunately, this system of equations is underconstrained and classical information theory offers no axioms to uniquely determine the desired decomposition. Something genuinely novel has to be added to the theory. This, in essence, is called the *PID problem*. Over the past decade, a range of proposals for concrete solutions have been presented in the literature (see Chapter 5.4, Figure 5.6). However, thus far no consensus could be reached.

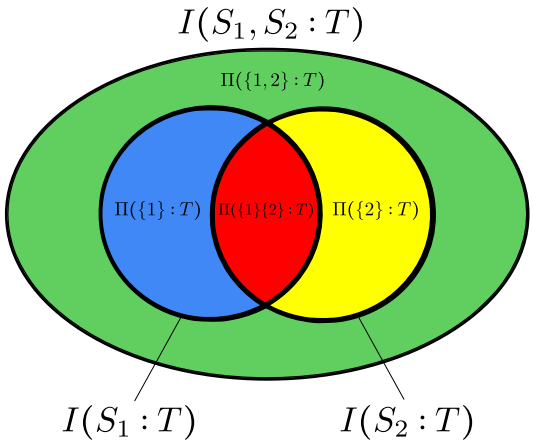

Fig. 1.1: Information diagram depicting the partial information decomposition for the case of two information sources. The inner two black circles represent the mutual information provided by the first source (left) and the second source (right) about the target. Each of these mutual information terms contains two atomic parts: $I(S_1 : T)$ consists of the unique information in source 1 ($\Pi(\{1\} : T)$, blue patch) and the information shared with source 2 ($\Pi(\{1\}\{2\} : T)$, red patch). $I(S_2 : T)$ consists of the unique information in source 2 ($\Pi(\{2\} : T)$, yellow patch) and again the shared information. The joint mutual information $I(S_1, S_2 : T)$ is depicted by the large black oval encompassing the inner two circles. $I(S_1, S_2 : T)$ consists of four atoms: The unique information in source 1 ($\Pi(\{1\} : T)$, blue patch), the unique information in source 2 ($\Pi(\{2\} : T)$, yellow patch), the shared information ($\Pi(\{1\}\{2\} : T)$, red patch), and additionally the synergistic information ($\Pi(\{1, 2\} : T)$, green patch).

In the general n-sources case, the problem becomes even more pronounced because the number of information atoms grows much more quickly than the number of constraints provided by classical information theory. Here the decomposition leads to information atoms $\Pi(\mathbf{a}_1, \ldots, \mathbf{a}_m : T)$ quantifying the information about the

target that can be obtained if and only if at least one of the source collections $\mathbf{a}_i \subseteq \{S_1, \ldots, S_n\}$ is known. There is one such atom per set of source collections $\alpha = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ such that no $\mathbf{a}_i$ is a subset of some distinct $\mathbf{a}_j$. Such collections α are called *antichains* and their number for a given n is known as the n-th Dedekind number.

The Dedekind numbers are a super-exponentially growing and very difficult to compute sequence of numbers. In fact, the ninth Dedekind number, a number with 42 digits, has been computed just half a year prior to the submission of this thesis by two independent research teams [20, 21]. By contrast, the number of constraints we obtain from classical information theory grows only exponentially: for n sources there are 2^{n-1} mutual information terms that we may relate to the information atoms in order to constrain them.

The exact construction of the general information decomposition from first principles, and different ways to obtain a specific solution, is the topic of Chapters 3-5. For now, let us turn our attention to the scientific relevance of the PID problem. Why is the PID problem important? Why would it be useful to have a solution to it?

1.3.2 Why care about PID?

PID illuminates existing information quantities First, PID can help clarify and provide better intuition for classical information theoretic quantities. Using the chain rule of mutual information the relations 1.24-1.26 can be written equivalently in terms of conditional mutual information

$$I(S_1 : T|S_2) = \Pi(\{1\} : T) + \Pi(\{1,2\} : T) \tag{1.27}$$

$$I(S_2 : T|S_1) = \Pi(\{2\} : T) + \Pi(\{1,2\} : T) \tag{1.28}$$

In other words, once we already know one of the sources, the additional information provided by the other should consist of the unique information it carries about the target plus the synergy of the two sources. This helps explain, for example, why conditional mutual information may be preferable to mutual information as a feature selection criterion in machine learning [22]. Ideally, a new feature should contain novel information given the already chosen features. This means that it should either contain some unique information about the correct output or contribute some information in a synergistic fashion together with the already chosen features. But according to the PID formalism, these are exactly the components provided by conditional mutual information. On the other hand, mutual information would

neglect the synergistic component and instead include information that is already redundantly contained in other features.

A second example is the interaction information (see Eq. 1.15) which has widely been considered as an indicator of redundancy (when negative) or synergy (when positive). The PID formalism can provide better insight into the correct interpretation. Using the relations above we find that

$$I(S_1 : S_2 : T) = I(S_1 : T) - I(S_1 : T|S_2) \tag{1.29}$$
$$= \Pi(\{1\} : T) + \Pi(\{1\}\{2\} : T) - (\Pi(\{1\} : T) + \Pi(\{1, 2\} : T)) \tag{1.30}$$
$$= \Pi(\{1\}\{2\} : T) - \Pi(\{1, 2\} : T) \tag{1.31}$$

So interaction information should be understood as the *difference* between redundancy and synergy. Assuming $\Pi(\{1\} : T)$ and $\Pi(\{1, 2\} : T)$ to be non-negative, this further implies bounds on redundancy and synergy

$$\Pi(\{1\}\{2\} : T) \geqslant I(S_1 : S_2 : T) \text{ if } I(S_1 : S_2 : T) > 0 \tag{1.32}$$
$$\Pi(\{1, 2\} : T) \geqslant -I(S_1 : S_2 : T) \text{ if } I(S_1 : S_2 : T) < 0 \tag{1.33}$$

Finally, PID also sheds light on the Wiener-Granger conception of causality and its non-parametric implementation via Transfer Entropy. Recall that Granger explicitly used the term "unique information" when justifying his choice of the term "causality". Now from a PID perspective this can be further refined. The additional information that the history of a source process $\mathbf{Y}_t$ provides about a target process $\mathbf{X}_t$ over and above the information in its own past is not just what is unqiuely contained in the history of $\mathbf{X}_t$. Including the past of the source process as a predictor might also make some synergistic information available, i.e. information we only obtain about $\mathbf{X}_t$ once we know *both* the target's and the sources' histories. Formally, we can again apply the fundamental PID equation for conditional mutual information here (Eq. 1.27), using $T := \mathbf{X}_{t+1}$ as the PID target and $S_1 := \mathbf{X}_t^-$ and $S_2 := \mathbf{Y}_t^-$ as PID sources:

$$TE_{\mathbf{Y}\to\mathbf{X}} = I(\mathbf{X}_{t+1} : \mathbf{Y}_t^- | \mathbf{X}_t^-) = \Pi(\{2\} : T) + \Pi(\{1, 2\} : T) \tag{1.34}$$

The unique information component is often called "state-independent" transfer entropy while the synergistic component is called "state-dependent" transfer entropy [23].

PID helps to formalize key concepts in the study of complex systems Complex systems such as neural networks are more than just "systems with many components". Even though the precise characterization and delineation of what it means for a system to be "complex" is a contentious area, one common theme is that, in some sense, a complex system is "more than just the sum of its parts". In other words, complex systems have *emergent* properties at the macroscopic level that are not easily explainable by just looking at the dynamics of its microscopic components [24] or that in some sense have a "life of their own" [25].

Even though the notion of emergence is a popular one, it is not easy to formalize mathematically. Here PID theory may offer a solution via the notion of synergistic information. Just like emergence, the concept of synergy is tightly connected to the idea of "the whole being more than the sum of its parts". Rosas et al. [26] introduced an information theoretic condition of what they call "causal emergence" based on PID ("causal" is here understood in the predictive Wiener-Granger sense). It is framed in terms of the predictability of the future system behaviour on the basis of smaller subsystems. Specifically, if there is some information about the future state of the system that can only be obtained if the states of more than k system components are known at the same time, then the system is said to exhibit emergence at order k. In the most extreme case (where k equals the number of system components), this is the information about the future state that can only be obtained by observing all the system components at the same time.

Some have suggested that it is a useful perspective on at least some complex systems, and in particular the brain, to think of them as *computational systems*, i.e. systems that compute some output (e.g. motor behaviour) based on some input (sensory experience) and the internal state of the system. Computational systems can be described via three fundamental operations: information storage, transfer, and modification [27, 28]. As discussed, in Section 1.2 above, for the first two operations there are already widely accepted measures based on classical information theory. However, the same is not the case for the modification of information. PID has been suggested as a useful framework to solve this problem. In particular, Lizier et al. [29] propose to measure information modification as the synergistic part of information transfer (see Equation 1.34 above). Here the underlying idea is that this part arises through some form of interaction with information stored in the target. It is this process of interaction that Lizier et al. conceptualize as modifying the information coming from the source.

Even the notion of complexity itself has been addressed utilizing the PID formalism. This is perhaps not surprising given the tight connection of this notion to emergence,

and hence synergistic information. Ehrlich et al. [30] introduced a measure of the *representational complexity* of neural networks. Consider a neural network with nodes $\mathbf{S} = (S_1, \ldots, S_n)$ aiming to represent some aspect of the external world T, e.g. "is there a frog over there or not?". Now the question is how much of the network we have to know in order to decode this information. Lets consider the extreme cases: on one end of the spectrum there might be a single node in the network from which the relevant aspect can be read of. In this case one might say that the representation is very simple. On the other end of the spectrum, it might be necessary to observe the entire network to obtain the desired information about the world while no proper subset of neurons tells us anything. In this case the representation can be considered to be very complex. It is captured by the complex multivariate relationships within the network.

The notion of representational complexity makes this intuition formal using PID theory. It is given by the average "degree of synergy" of the information atoms making up the total mutual information $I(S_1, \ldots, S_n : T)$ the network carries about T. The degree of synergy, denoted by $m(\alpha)$, is the minimal number of source variables one needs in order to obtain the information atom $\Pi(\alpha : T)$ (recall that α is an antichain of source collections). Accordingly, it is a natural number between 1 and n. Each possible degree of synergy is now weighted by the percentage of the total mutual information provided by information atoms of that degree. Formally, the representational complexity $C(\mathbf{S} : T)$ is then defined as

$$W_k := \frac{\sum_{m(\alpha)=k} \Pi(\alpha : T)}{I(\mathbf{S} : T)}, \tag{1.35}$$

$$C(\mathbf{S} : T) := \sum_{k=1}^{n} k W_k. \tag{1.36}$$

In other words, for each $1 \leqslant k \leqslant n$ we may ask: What percentage of the total information about T can we obtain by looking at subsets of exactly k neurons? Then we weight k by that percentage. The result represents something like the "average number of network nodes we need to observe in order to decode the target". In the extreme cases described above we obtain $C = 1$ and $C = n$, respectively.

PID can be used to test and construct theories of (neural) processing One promising application of PID lies in its capacity for testing theories of neural processing based on their distinct information-theoretic implications. Different theories have, as it were, different 'information-theoretic footprints' [31]. For example, certain theories may propose the existence of neurons that serve informationally distinct

roles within a network. A case in point is predictive processing theories, which distinguish between 'error units' and 'representation units,' each tasked with handling different types of information [32]. Another example, from machine learning, is the information bottleneck theory [33] that suggests two distinct phases in the learning process of artificial neural networks: an error-minimization phase during which the network seeks to improve its prediction of the correct output as much as possible and a compression phase in which the network attempts to form a maximally simple representation of the input. Each phase is characterized by specific behaviour the mutual information between the hidden layers and the output or the input, respectively.

PID might be fruitfully used in this context because it affords a particularly detailed information theoretic footprint: it is able to describe all the multitude of possible informational relationships between variables. A particularly promising way to carry out this approach is to directly phrase theories in information theoretic terms. This can be achieved for instance via the framework of *information theoretic goal-functions* [31] leading to so called *infomorphic networks* [34]. The starting point is a neuron with two distinct types of input: receptive input $\mathbf{S}_R$ from hierarchically lower layers of the network and contextual input $\mathbf{S}_C$ from higher layers. This sort of distinction is for example very useful in describing layer-5 pyramidal neurons which have two different types of dendrites. The basal dendrites which are thought to mediate the perceptual input and the apical dendrites responsible for the contextual input. The information in the output T of the neuron, i.e. its entropy $H(T)$, can be decomposed as

$$H(T) = I(\mathbf{S}_R, \mathbf{S_C}\colon T) + H(T|\mathbf{S}_R, \mathbf{S_C}) \tag{1.37}$$

$$= \Pi(\{\mathbf{S}_R\}\{\mathbf{S}_C\}\colon T) + \Pi(\{\mathbf{S}_R\}\colon T) + \Pi(\{\mathbf{S}_C\}\colon T) + \Pi(\{\mathbf{S}_R\mathbf{S}_C\}\colon T) + H(T|\mathbf{S}_R, \mathbf{S_C}) \tag{1.38}$$

An information theoretic goal function is now constructed by giving each term in the sum a weight Γ_i. Different choices of these weights correspond to different goal functions leading the neural processor to prioritize certain types of information over others. For example, it might place a lot of weight on redundant information that is contained both in the receptive input and the contextual input, and is in this way supported not only from sensory data but also from information already available to the system. Or it might want to minimize the information only contained in the contextual input while at the same time allowing some information to be uniquely provided by the receptive input. In this way a very broad class of goal functions is obtained encompassing some that had previously been suggested in the literature (e.g. "coherent infomax") and also some that were originally not phrased in information theoretic terms (e.g. certain forms of predictive coding). It has been demonstrated that this approach can usefully be employed to various

learning paradigms such as supervised learning, unsupervised learning, as well as associative memory learning [34]. It can also be extended to include more types of inputs and more complicated network structures. One interesting question in this line of research would be whether real neurons can be parsimoniously described by a small set of information theoretic goal functions.

1.3.3 Contribution of this work

Chapters 3-5 are concerned with the mathematical *structure* of the PID problem as well as the development of a concrete solution, i.e. particular *measures* of the different PID atoms.

Chapter 3 ("Introducing a differentiable measure of pointwise shared information") addresses the later aspect by proposing a general measure of redundant information called the shared exclusions measure i^{sx}. It was already shown in the original exposition of PID theory by Williams and Beer that, even in the general n-sources case, a unique solution for all information atoms can be obtained once a measure of the redundant information of arbitrary source collections is specified. The derivation of the shared exclusions measure closely follows the construction of mutual information, thus preserving many basic information theoretic intuitions. It is formulated on the *pointwise* level of individual realizations of source and target variables. This perspective was used by Fano [35] as a starting point from which the entirety of information theory can be derived. Due to the similarity in construction i^{sx} inherits some very useful properties from pointwise mutual information, in particular its continuity and differentiability with respect to the underlying joint distribution and also a target chain rule. These properties make it especially useful for applications in the context of artificial neural networks where it can be used to formulate information theoretic goal functions [31] as discussed in the previous section. The chapter provides two distinct ways to motivate the shared exclusions measure, establishes its mathematical properties and operational interpretation, and illustrates the entailed decomposition for some exemplary probability distributions.

Chapters 4-5 are concerned with the mathematical structure of the PID problem. Chapter 4 ("Bits and pieces: understanding information decomposition from part-whole relations and formal logic") shows that the entire theory of PID can be derived, firstly, from considerations of part-whole relationships between information atoms and mutual information terms, and secondly, based on a hierarchy of logical constraints describing how a given information atom can be accessed. In this way, the idea of a PID is developed on the basis of two of the most elementary relationships

in nature: the part-whole relationship and the relation of logical implication. This unifying perspective provides insights into pressing questions in the field such as the possibility of constructing a PID based on concepts other than redundant information in the general n-sources case. The paper also presents a re-derivation of the shared exclusions measure of redundant information introduced in Chapter 3 based on principles of logic and mereology (the study of part-whole relationships).

Chapter 5 ("From Babel to Boole: The Logical Organization of Information Decompositions") centers around the notion of PID *base-concepts*, i.e. information functionals that induce an entire information decomposition once they are defined in terms of the underlying joint probability distribution of source and target variables. The standard functional used as a PID base-concept is redundant information, yet there has been ongoing interest in examining the problem through the lens of different base-concepts of information, such as synergy, unique information, or union information. The parthood formulation of PID which was introduced in "Bits and Pieces" showed that PID base-concepts can be expressed in terms of conditions phrased in formal logic on the specific parthood relations between the PID components and the different mutual information terms. "From Babel to Boole" builds on this foundation by setting forth a general pattern of these logical conditions of which all PID base-concepts in the literature are special cases and that also reveals novel base-concepts, in particular a concept we call “vulnerable information”.

1.4 Graph statistics

Graph theory often enters into consideration naturally after an information-theoretic analysis. This is because information theoretic measures are frequently used to evaluate the informational relationships between any pair of nodes in a network, and these relationships can naturally be represented as a graph for further analysis. In a neuroscientific context, the networks would typically consist of certain brain regions of interest, the activity of which might be measured using various neuroimaging techniques.

The graphs obtained in this way may be undirected, for instance if mutual information is used as a measure of connectivity, or they could be directed. A typical example of the latter type would be Transfer Entropy based network inference as described in [36] (see Fig.1.2 for an illustration). Here, each network node is considered in turn as a target node. The goal is to identify a maximally informative set of source nodes, but in such a way that no superfluous sources are included.

This is achieved by requiring that each source should provide significant Transfer Entropy about the target conditional on all other nodes in the set. Another example of a directed information theoretic analysis would be a pairwise Granger causality analysis where the Granger causalities from any node in the network to any other node are estimated conditional on the rest of the network (see e.g. [37]).

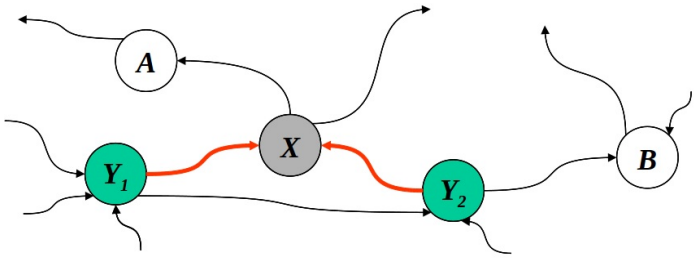

Fig. 1.2: Illustration of the main idea behind transfer entropy based network inference. The goal is to infer, for each node in the network (for instance X), the set of source nodes (here Y_1 , Y_2) from which from information is transferred into the given node. The Figure was adopted from Lizier and Rubinov, 2012

Quite often, networks of informational relationships (or "functional connectivity") are estimated for experimental units in different groups or under different experimental conditions with the aim to find differences in these relationships between the groups/conditions. The process of inferring these functional networks can be though of as a measurement of two *graph generating processes*, i.e. processes that randomly, accordingly to a certain probability distribution, generate graphs describing the informational relationships between network nodes. Suppose for example that we are measuring two groups, a group of patients with Asperger Spectrum Disorder (ASD) and a control group. Then we might be interested in the question: is the pattern of information flow between certain brain regions of interest different (on average) in ASD patients versus subjects in the control group? Maybe the information flow between two regions is more pronounced in one group, or it occurs at a different temporal delay, or it always goes along with information flow to a third region while it does not do so in the other group.

In order to address questions of this nature, we need statistical methods for testing differences between the graph generating processes. Importantly, due to the immense number of possible differences, these methods have to efficiently control for multiple comparisons in order to guarantee bounded false positive rates.

1.4.1 Contribution of this work

Chapter 6 ("Significant Subgraph Mining for Neural Network Comparison with Multiple Comparisons Correction") describes how the method of Significant Subgraph Mining can be fruitfully employed in the context of neural network comparison. Significant Subgraph Mining is a recently developed [38, 39] method for statistically comparing processes generating binary graphs, i.e. an edge can be present or absent but has no numerical value attached to it.

The project developed exactly in the context alluded to above, i.e. after a Transfer Entropy based network analysis of 20 ASD patients and 20 neurotypical controls using resting state MEG recordings. This analysis lead to a binary, directed graph for each subject where the edges indicate whether there was statistically significant information flow from one region of interest to another. Additionally, each edge is labelled with possibly multiple time lags, indicating at what temporal delays the information flow was detected. The seven regions of interest used in the analysis had already been suggested by a previous study [40]. Now, the question arose how to systematically look for differences between the graphs in the two groups. There was no theoretical reason to restrict the search to differences of a particular type such as differences in the occurrences of individual edges (e.g. "significant information flow from region A to region B occurred with a higher probability in the ASD group"). Rather, a more general pattern mining approach that systematically looks for all possible differences between the graph generating processes (be they at the level of edges, or dependencies between edges, or even more complex higher order interactions) seemed in order.

In principle, any possible stochastic difference between the two graph-generating processes can be expressed in terms of the probabilities of occurrence of specific subgraphs. Significant Subgraph Mining systematically tests all such differences while correcting for the formidable multiple comparisons problem arising because the total number of possible subgraphs scales super-exponentially in the number of graph nodes. However, the original subgraph mining method was in some respects not adapted to circumstances frequently arising in neuroscience research. This gap is addressed in Chapter 6.

In particular, it extends the method from between-subject designs (i.e. independent graph-generating processes) to within-subject experimental designs that allow for dependencies between the graph-generating processes. It also provides a systematic analysis of its error-statistical properties in simulation using Erdős-Rényi models and it presents an empirical power analysis utilizing the MEG data set mentioned

above. Based on these analyses, practical recommendations for the application of subgraph mining in neuroscience are derived. Finally, a python implementation as part of the openly available IDTxl toolbox is provided. This implementation directly takes account of the data structures arising in information theoretic analyses of neuroimaging data, e.g. the possibility for information transfer to occur at different time delays. In this way the adapted method is also able to detect differences with respect to the temporal structure of information flow.

2

Sampling distribution for single-regression Granger causality estimators

Aaron J. Gutknecht [1], Lionel Barnett [2],

1 Campus Institute for Dynamics of Biological Networks, Georg-August University, Goettingen, Germany

2 Sussex Centre for Consciousness Science, Department of Informatics, University of Sussex, Falmer, Brighton BN1 9RH, U.K.

Published as: ***A J Gutknecht, L Barnett, Sampling distribution for single-regression Granger causality estimators, Biometrika, Volume 110, Issue 4, December 2023, Pages 933–952, https://doi.org/10.1093/biomet/asad009***

Abstract

The single-regression Granger-Geweke causality estimator has previously been shown to solve known problems associated with the more conventional likelihood-ratio estimator; however, its sampling distribution has remained unknown. We show that, under the null hypothesis of vanishing Granger causality, the single-regression estimator converges to a generalized χ^2 distribution, which is well approximated by a Γ distribution. We show that this holds too for Geweke's spectral causality averaged over a given frequency band, and derive explicit expressions for the generalized χ^2 and Γ-approximation parameters in both cases. We present a Neyman–Pearson test based on the single-regression estimators, and discuss how it may be deployed in empirical scenarios. We outline how our analysis may be extended to the conditional case, point-frequency spectral Granger causality, and the important case of state-space Granger causality.

2.1 Introduction

Since its inception in the 1960s, Wiener-Granger causality has found many applications in a range of disciplines, from econometrics, neuroscience, climatology, ecology, and beyond. In the early 1980s Geweke introduced the standard vector-autoregressive (VAR) formalism, and the Granger-Geweke population loglikelihood-ratio statistic [10, 11]. As well as furnishing a likelihood-ratio test for statistical significance, the statistic has been shown to have an intuitive information-theoretic interpretation as a quantitative measure of information transfer between stochastic processes [13, 41]. In finite sample, the likelihood-ratio estimator requires separate estimates for the full and reduced VAR models, and as such admits the classical large-sample theory, and asymptotic χ^2 distribution [42–44]. However, it has become increasingly clear that the "dual-regression" likelihood-ratio estimator is problematic: specifically, model order selection involves a bias-variance trade-off which may potentially lead to spurious results, including negative Granger causality values [14, 45, 46].

More recently, an alternative *single-regression* estimator which obviates the problem has been developed in various forms [18, 47–49]; but, since the large-sample theory no longer obtains, its sampling distribution has thus far remained unknown. In addition to the reduced bias and variance of the single-regression estimator [49], knowledge of its sampling distribution under the null hypothesis of vanishing causality would allow to construct novel hypothesis tests, especially in the frequency domain where little is known about the sampling distribution of Geweke's spectral Granger causality statistic [10, 11]. Closing this gap is the central objective of the present study. Our novel application of the 2nd-order Delta Method [50], furthermore, opens a path to significant extensions of our result, in particular to the sampling distribution of the state-space Granger causality estimator [18, 51], which remains unknown.

2.2 VAR modelling

We assume given a wide-sense stationary, purely-nondeterministic n-dimensional vector stochastic process $U_t = [U_{1t}, \ldots, U_{nt}]^\mathsf{T}$, $-\infty < t < \infty$. U_t then has a Wold

moving-average decomposition [52], which we assume may be inverted to yield a stable (in general infinite-order) VAR representation

$$U_t = \sum_{k=1}^{\infty} A_k U_{t-k} + \varepsilon_t \,, \tag{2.1}$$

where ε_t is a white noise innovations process. The sequence of $n \times n$ autoregression coefficient matrices A_k is square-summable, and, since the process is purely-nondeterministic, the $n \times n$ residuals covariance matrix $\Sigma = \mathrm{E}[\varepsilon_t \varepsilon_t^\mathsf{T}]$ is positive-definite. Following Geweke [10], we further assume that the cross-power spectral density (CPSD) matrix for U_t is uniformly bounded away from zero; this guarantees a stable, invertible VAR representation for any *sub*process of U_t [53, 54]. We assume these conditions for all vector stochastic processes from now on.

If $A_k = 0$ for $k > p$ then (2.1) defines a finite-order VAR(p) model. We write $A = [A_1 \dots A_p]$ (an $n \times pn$ matrix), and the model parameters are $\theta = (A, \Sigma)$. The autocovariance sequence for the process U_t is $\Gamma_k = \mathrm{E}\big[U_t U_{t-k}^\mathsf{T}\big]$ ($-\infty < k < \infty$), and $\Gamma_{-k} = \Gamma_k^\mathsf{T}$. By a standard trick, the process $\tilde{U}_t = \big[U_t^\mathsf{T}\ U_{t-1}^\mathsf{T} \dots U_{t-p+1}^\mathsf{T}\big]^\mathsf{T}$ satisfies the pn-dimensional VAR(1) model

$$\tilde{U}_t = \tilde{A}\tilde{U}_{t-1} + \tilde{\varepsilon}_t \,, \tag{2.2}$$

where the $pn \times pn$ VAR "companion matrix" $\tilde{A}$ and covariance matrix $\tilde{\Sigma}$ of the residuals $\tilde{\varepsilon}_t$ are respectively

$$\tilde{A} = \begin{bmatrix} A_1 & A_2 & \dots & A_{p-1} & A_p \\ I & 0 & \dots & 0 & 0 \\ 0 & I & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & I & 0 \end{bmatrix}, \qquad \tilde{\Sigma} = \begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}.$$

The *spectral radius* of the model is defined as the largest absolute eigenvalue of $\tilde{A}$:

$$\rho(A) = \max\{|z| : \big|Iz - \tilde{A}\big| = 0\} \,. \tag{2.3}$$

The model is stable iff $\rho(A) < 1$.

Taking the covariance of both sides of (2.2) yields

$$\tilde{\Gamma} = \tilde{A}\tilde{\Gamma}\tilde{A}^\mathsf{T} + \tilde{\Sigma} \,, \tag{2.4}$$

where $\tilde{\Gamma}$ is the $pn \times pn$ block-Toeplitz covariance matrix

$$\tilde{\Gamma} = \mathrm{E}\left[\tilde{U}_t \tilde{U}_t^{\mathsf{T}}\right]. \tag{2.5}$$

The $k\ell$-block of $\tilde{\Gamma}$ is $\tilde{\Gamma}_{k\ell} = \Gamma_{\ell-k}$ $(k, \ell = 1, \ldots, p)$. Eq. (2.4) is a discrete-time Lyapunov (DLYAP) equation, which may be readily solved numerically. If the parameters (A, Σ) are known, the Γ_k may be calculated from (2.4); conversely, (A, Σ) may be calculated from the Γ_k, e.g., by Whittle's algorithm [55]. In sample, maximum-likelihood parameter estimates $\left(\hat{A}, \hat{\Sigma}\right)$ may be calculated via a standard ordinary least squares (strictly speaking, OLS only yields true maximum-likelihood estimates in the case that the innovations are multivariate-normal).

In the spectral domain [56], let $\omega \in [0, 2\pi]$ denote angular frequency in radians. The *transfer function* $\Psi(\omega)$ for the VAR model (2.1) is defined as

$$\Psi(\omega) = \Phi(\omega)^{-1}, \qquad \Phi(\omega) = I - \sum_{k=1}^{\infty} A_k e^{-i\omega k}. \tag{2.6}$$

The cross-power spectral density matrix $S(\omega)$ is the Fourier transform of the autocovariance sequence and, conversely, the autocovariance sequence is the inverse transform of the CPSD:

$$S(\omega) = \sum_{k=-\infty}^{\infty} \Gamma_k e^{-i\omega k}, \qquad \Gamma_k = \frac{1}{2\pi} \int_0^{2\pi} S(\omega) e^{i\omega k} d\omega. \tag{2.7}$$

The matrix $S(\omega)$ is Hermitian for all ω, and satisfies the factorisation [57]

$$S(\omega) = \Psi(\omega) \Sigma \Psi(\omega)^*. \tag{2.8}$$

The CPSD uniquely determines the VAR parameters, and computationally $\Psi(\omega)$ and Σ may be factored out from (2.8), e.g., by Wilson's algorithm [58].

2.3 Granger-Geweke causality

2.3.1 The population statistic

Geweke [10] defines the population (unconditional) Granger causality statistic in the following context: suppose that the process (2.1) is partitioned into subprocesses $U_t = [X_t^{\mathsf{T}}\ Y_t^{\mathsf{T}}]^{\mathsf{T}}$ of dimension n_x, n_y respectively. The assumed regularity conditions

on U_t [10, Sec. 2] ensure that the subprocess X_t will itself admit a stable, invertible VAR representation ("reduced regression")

$$X_t = \sum_{k=1}^{\infty} A_k^{\mathsf{R}} X_{t-k} + \varepsilon_t^{\mathsf{R}} \tag{2.9}$$

with square-summable coefficients A_k^{R} and positive-definite residuals covariance matrix $\Sigma^{\mathsf{R}} = \mathrm{E}[\varepsilon_t^{\mathsf{R}} \varepsilon_t^{\mathsf{R}\mathsf{T}}]$ (superscript 'R' will be used generally to refer to quantities associated with the reduced regression). To define Granger-Geweke causality, the prediction error of the reduced regression (2.9) is contrasted with that of the "full regression"; that is, the x-component

$$X_t = \sum_{k=1}^{\infty} A_{k,xx} X_{t-k} + \sum_{k=1}^{\infty} A_{k,xy} Y_{t-k} + \varepsilon_{xt} \tag{2.10}$$

of (2.1). We stress that (i) the reduced model parameters $(A^{\mathsf{R}}, \Sigma^{\mathsf{R}})$ are fully determined by the full model parameters (A, Σ), and (ii) even if the full regression (2.10) has finite order, the reduced regression (2.9) will in general *not* have finite order. The Granger-Geweke causality measure (henceforth just "Granger casuality") from $Y \to X$ for the VAR model (2.1) with parameters θ is then defined as

$$F_{Y \to X}(\theta) = \log \left|\Sigma^{\mathsf{R}}\right| - \log |\Sigma_{xx}|. \tag{2.11}$$

Intuitively, it measures the degree to which the (linear least-squares) prediction of X can be improved by taking into account the past of Y, as compared with prediction of X based only on its own past. It may also be interpreted as an approximation to the "information transfer" from Y to X, on the basis that under Gaussian assumptions the Granger causality statistic (2.11) is asymptotically equivalent to the more general non-parametric transfer entropy [6, 13]

In the frequency domain, Geweke [10] defines the (population, unconditional) spectral Granger causality measure at angular frequency ω by

$$f_{Y \to X}(\omega; \theta) = \log |S_{xx}(\omega)| - \log \left|S_{xx}(\omega) - \Psi_{xy}(\omega) \Sigma_{yy|x} \Psi_{xy}(\omega)^*\right|, \tag{2.12}$$

where

$$\Sigma_{yy|x} = \Sigma_{yy} - \Sigma_{yx} [\Sigma_{xx}]^{-1} \Sigma_{xy} \tag{2.13}$$

is a partial covariance matrix. Spectral Granger causality addresses the extent to which variance of X may be explained by variance of Y at a given frequency ω; e.g., "in the long run" for low frequencies, or "in the short run" for high frequencies. For a concrete econometric example see Geweke [10].

Barnett and Seth [59] introduce *band-limited* (frequency-averaged) spectral Granger causality

$$f_{Y\to X}(\mathcal{F};\theta) = \frac{1}{|\mathcal{F}|}\int_{\mathcal{F}} f_{Y\to X}(\omega;\theta)\,d\omega\,, \tag{2.14}$$

where the frequency range $\mathcal{F}$ is a measurable subset of $[0, 2\pi]$, in practice usually an interval. Averaging $f_{Y\to X}(\omega;\theta)$ across all frequencies, we recover the corresponding time-domain Granger causality; that is [10],

$$f_{Y\to X}([0,2\pi];\theta) = \frac{1}{2\pi}\int_0^{2\pi} f_{Y\to X}(\omega;\theta)\,d\omega = F_{Y\to X}(\theta)\,. \tag{2.15}$$

The band-limited statistic is of particular interest in neuroscience applications, since functional and cognitive phenomena in neural systems are well-known to be strongly associated with spectral power in specific frequency bands [60].

2.3.2 Likelihood-ratio estimation

Suppose given a finite-order VAR model

$$U_t = \sum_{k=1}^{p} A_k U_{t-k} + \varepsilon_t \tag{2.16}$$

for the process $U_t = [X_t^\mathsf{T}\ Y_t^\mathsf{T}]^\mathsf{T}$. For now, we assume that the model order p is known (in Section 2.5.4 we discuss infinite-order VAR models and model order selection). In what follows, we write $\hat{\theta} = \big(\hat{A}, \hat{\Sigma}\big)$ to denote the random variable $\hat{\theta}(U)$, where $\hat{\theta}(u)$ is the maximum-likelihood estimate of the parameter θ for given time-series data u, and U is a stochastic process distributed according to the VAR model (2.16).

On the face of it, $F_{Y\to X}(\theta)$ is a population likelihood-ratio statistic [10], since the maximum likelihood (throughout, "likelihood" refers to *average loglikelihood*) for a finite-order VAR model of the form (2.16) is, up to an additive constant, $-\frac{1}{2}\log|\hat{\Sigma}|$, where $\hat{\Sigma}$ is the maximum-likelihood estimate for the population residuals covariance matrix Σ. As regards estimation of $F_{Y\to X}(\theta)$, however, while the full model order p may be finite, the reduced model (2.9) will in general be of *infinite* order. We might be tempted, as suggested in Geweke [10, 11], and until recently standard practice, to simply truncate the reduced model at order p. Then (2.9) becomes a nested sub-model of (2.10) corresponding to the null hypothesis of vanishing Granger causality:

$$H_0 : A_{1,xy} = \ldots = A_{p,xy} = 0\,. \tag{2.17}$$

The likelihood-ratio Granger causality estimator is then

$$\hat{F}^{\mathrm{LR}}_{Y\to X} = \log\left|\hat{\Sigma}^{\mathrm{R}}\right| - \log\left|\hat{\Sigma}_{xx}\right|, \tag{2.18}$$

where $\hat{\Sigma}^{\mathrm{R}}$ is the maximum-likelihood estimator for Σ^{R}; i.e., based on the reduced model (2.9). Note that the distribution of the estimator $\hat{F}^{\mathrm{LR}}_{Y\to X}$ depends on the actual parameters θ.

Here, though, a problem arises: in general, the truncated reduced model will be misspecified, and failure to take into account sufficient lags of X_t in the reduced regression biases the estimator (2.18). Noting that a VAR(p) model is also VAR(q) for $q > p$, we could attempt to remedy the situation by selecting a parsimonious model order $q > p$ for the reduced model by a standard model order selection criterion [61], and extend the full model to order q. However, in doing so the full model becomes over-specified and the variance of the resulting estimator is inflated. Furthermore, since the estimated model order will generally increase with sample length N, it is not clear whether the estimator will be consistent in any meaningful sense. We discuss this further in Section 2.5.4. This conundrum was explicitly identified by Stokes and Purdon [14], although its symptoms had previously been noted, particularly in the spectral domain [see e.g., 45, 46]. We remark that Stokes and Purdon [14], having identified the likelihood-ratio estimator as problematic, concede that at the time they were unaware that there were already estimators which obviate the problem [17]. For further commentary on the issues raised in Stokes and Purdon [14], see Barnett et al. [15, 16], Faes et al. [62], and Dhamala et al. [63]. We note also that the "block-decomposition" method presented in Chen et al. [46]—essentially an attempt at constructing a single-regression estimator (see Section 2.3.3 below)—is incorrect [51].

As a likelihood-ratio statistic, $\hat{F}^{\mathrm{LR}}_{Y\to X}$ obeys Wilks' Theorem [43], which implies that for any $\theta \in \Theta_0$

$$N\hat{F}^{\mathrm{LR}}_{Y\to X} \xrightarrow{d} \chi^2(d) \tag{2.19}$$

as sample size $N \longrightarrow \infty$, with degrees of freedom $d = qn_xn_y$, where q is the selected model order for the full and reduced models. Convergence is of order $N^{-1/2}$. Note that it should not be assumed that the bias/variance trade-off discussed above is necessarily problematic as regards statistical inference; see Section 2.5.2 below.

2.3.3 Single-regression estimation

The above problem may be sidestepped. Given a finite-order VAR(p) model (2.16) (again, we assume that p is known), the reduced VAR (2.9) may not be assumed finite-dimensional, but the reduced residuals covariance matrix Σ^{R} will, as previously remarked, be a continuous, *deterministic* function

$$\Sigma^{\mathrm{R}} = V(\theta) \tag{2.20}$$

of the finite-dimensional full-model parameters $\theta = (A, \Sigma)$, with $V(\theta) = \Sigma_{xx}$ for $\theta \in \Theta_0$. Given parameters (A, Σ), the function $V(\theta)$ may be computed numerically to desired precision by spectral factorisation in the frequency domain [47, 48], spectral factorisation in the time domain [49], a linear transformation/autocovariance method due to Dufour and Taamouti [64], or by a state-space method [18, 51] which devolves to solution of a discrete algebraic Riccati equation (DARE); see Supplementary Material, Section 2.7.3. From (2.11) and (2.20) the population Granger causality is

$$F_{Y\to X}(\theta) = \log|V(\theta)| - \log|\Sigma_{xx}|\,. \tag{2.21}$$

Given a data sample, we need only obtain the maximum-likelihood parameter estimate $\hat{\theta} = \left(\hat{A}, \hat{\Sigma}\right)$ for the full model (2.16); the estimate for Σ^{R} may then be calculated directly from $\hat{\theta}$ as $V\left(\hat{\theta}\right)$, by one of the methods mentioned above, yielding the *single-regression Granger causality estimator*

$$\hat{F}^{\mathrm{SR}}_{Y\to X} = \log\left|V\left(\hat{\theta}\right)\right| - \log\left|\hat{\Sigma}_{xx}\right|. \tag{2.22}$$

Since maximum-likelihood parameter estimates are consistent, $\hat{\Sigma}_{xx} \xrightarrow{p} \Sigma_{xx}$, and by (2.20) and the Continuous Mapping Theorem [CMT; 50] $V\left(\hat{\theta}\right) \xrightarrow{p} \Sigma^{\mathrm{R}}$. Thus the single-regression estimator (2.21) is a consistent estimator of $F_{Y\to X}(\theta)$.

The single-regression estimator (2.22) is not a likelihood ratio, so Wilks' Theorem does not apply [*cf.* (2.19)], and the asymptotic distribution under the null hypothesis (2.17) has thus far remained unknown. We shall see that, in contrast to Wilks' asymptotic χ^2 null distribution, the sampling distribution of the single-regression estimator under the null depends explicitly on the (true) null parameter $\theta \in \Theta_0$ itself. This raises some issues regarding statistical inference which we address in Section 2.5.

2.4 Asymptotic null distribution for single-regression estimators

2.4.1 The 2nd-order Delta Method

We proceed with a technical result, a multivariate 2nd-order Delta Method [50], on which our derivation of the asymptotic distributions for the time-domain and band-limited spectral estimators hinges:

Proposition 1. *Let $f(\theta)$ be a non-negative, twice-differentiable function on a smooth r-dimensional manifold $\Theta \subseteq \mathbb{R}^r$ which vanishes identically on the s-dimensional hyperplane $\Theta_0 \subset \Theta$ specified by $\theta_1 = \ldots = \theta_d = 0$ with $d = r - s$. Then*

a. The gradient $\nabla f(\theta)$ is zero for all $\theta \in \Theta_0$.

b. Writing a subscript "$_0$" to denote the $d \times d$ upper-left submatrix of an $r \times r$ matrix, for $\theta \in \Theta_0$ the Hessian $W(\theta) = \nabla^2 f(\theta)$ takes the form

$$W(\theta) = \begin{bmatrix} W_0(\theta) & 0 \\ 0 & 0 \end{bmatrix}$$

with $W_0(\theta)$ positive-semidefinite.

c. For $\theta \in \Theta_0$ let ϑ_N be a sequence of r-dimensional random vectors with $N^{1/2}(\vartheta_N - \theta) \xrightarrow{d} \mathcal{N}\big(0, \Omega(\theta)\big)$ as $N \longrightarrow \infty$. Then

$$N f(\vartheta_N) \xrightarrow{d} \chi^2\big(\tfrac{1}{2} W_0(\theta), \Omega_0(\theta)\big) \tag{2.23}$$

as $N \longrightarrow \infty$, where $\Omega_0(\theta)$ denotes the upper-left block of $\Omega(\theta)$, and $\chi^2(A, B)$ denotes the generalized χ^2 distribution *[Mohsenipour [65]; see Supplementary Material, Section 2.7.1].*

Proof. See Supplementary Material, Section 2.7.2. □

In the more general case where Θ_0 is a smooth submanifold of Θ rather than a simple hyperplane, under an appropriate local change of coordinates we again obtain a generalised χ^2 distribution of the form (2.23) (Supplementary Material, Section 2.7.2).

2.4.2 The time-domain single-regression estimator

We shall apply Proposition 1 to the $F_{Y\to X}(\theta)$ of (2.21). Firstly, $F_{Y\to X}(\theta)$ is non-negative and vanishes on Θ_0 [10]. Below we establish that it is also twice-differentiable (in fact real analytic). Secondly, by the large-sample theory, for any $\theta \in \Theta$ we have $\hat{\theta} \xrightarrow{p} \theta$, and $N^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Omega(\theta))$ as sample size $N \longrightarrow \infty$, where $\hat{\theta}$ is the maximum-likelihood parameter estimate, and $\Omega(\theta)$ the inverse of the *Fisher information matrix* for the VAR(p) model (2.16) evaluated at the parameter θ. Thus by Proposition 1, we have for any $\theta \in \Theta_0$

$$N\hat{F}^{\mathrm{SR}}_{Y\to X} \xrightarrow{d} \chi^2\left(\tfrac{1}{2}W_0(\theta), \Omega_0(\theta)\right) \tag{2.24}$$

as $N \longrightarrow \infty$, where $W(\theta)$ is the Hessian of $F_{Y\to X}(\theta)$. Here the "$_0$" subscript denotes a submatrix corresponding to the null-hypothesis variable indices $x = \{1, \ldots, n_x\}$, $y = \{n_x + 1, \ldots, n\}$.

To calculate the generalized χ^2 parameters, we thus require firstly the null submatrix $\Omega_0(\theta)$ of $\Omega(\theta)$. This is a standard result: let $\tilde{\Gamma}$ be the autocovariance matrix of (2.5). Considering multi-indices $[k, ij]$ for the regression coefficients $A_{k,ij}$ (so that k indexes lags and i, j variables), the entries for the inverse Fisher information matrix corresponding to the $A_{k,ij}$ are [66, 67]

$$\Omega(\theta)_{[k,ij][k',i'j']} = \Sigma_{ii'}\left[\tilde{\Gamma}^{-1}\right]_{kk',jj'} = \left[\Sigma \otimes \tilde{\Gamma}^{-1}\right]_{[k,ij][k',i'j']}, \tag{2.25}$$

where $\left[\tilde{\Gamma}^{-1}\right]_{kk',jj'}$ denotes the jj' entry of the kk'-block of $\tilde{\Gamma}^{-1}$, and "$\otimes$" the Kronecker matrix product. Then $\Omega_0(\theta)$ is the submatrix of (2.25) with $i, i' \in x$ and $j, j' \in y$, or

$$\Omega_0(\theta) = \Sigma_{xx} \otimes \left[\tilde{\Gamma}^{-1}\right]_{yy}. \tag{2.26}$$

Secondly, to calculate the null Hessian $W_0(\theta)$, we require an expression for the function $V(\theta)$ of (2.20). This we accomplish via the state-space formalism introduced in Barnett and Seth [18].

Proposition 2.

$$V(\theta) = A_{xy}\Pi A_{xy}^T + \Sigma_{xx}, \tag{2.27}$$

where the $pn_y \times pn_y$ symmetric matrix Π is the solution of the DARE

$$\Pi = \tilde{A}_{yy}\Pi\tilde{A}_{yy}^T + \tilde{\Sigma}_{yy} - \left(\tilde{A}_{yy}\Pi A_{xy}^T + \tilde{\Sigma}_{yx}\right)\left(A_{xy}\Pi A_{xy}^T + \Sigma_{xx}\right)^{-1}\left(\tilde{A}_{yy}\Pi A_{xy}^T + \tilde{\Sigma}_{yx}\right)^T \tag{2.28}$$

with

$$\tilde{A}_{yy} = \begin{bmatrix} A_{1,yy} & A_{2,yy} & \dots & A_{p-1,yy} & A_{p,yy} \\ I & 0 & \dots & 0 & 0 \\ 0 & I & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & I & 0 \end{bmatrix}, \quad A_{xy} = \begin{bmatrix} A_{1,xy} & A_{2,xy} & \dots & A_{p-1,xy} & A_{p,xy} \end{bmatrix},$$

and

$$\tilde{\Sigma}_{yy} = \begin{bmatrix} \Sigma_{yy} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}, \qquad \tilde{\Sigma}_{yx} = \begin{bmatrix} \Sigma_{yx} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Proof. See Supplementary Material, Section 2.7.3. □

It is not hard to see that, by construction, $F_{Y \to X}(\theta) = \log|V(\theta)| - \log|\Sigma_{xx}|$ is an analytic function of θ: calculation of $V(\theta)$ via (2.28) and (2.27) only involves algebraic operations (solution of multivariate polynomial equations), and we know $V(\theta)$ to be positive-definite, so that $|V(\theta)| > 0$ for all θ and $\log|V(\theta)|$ is thus analytic.

Our next result establishes an expression for $W_0(\theta)$.

Proposition 3.

$$W_0(\theta) = 2[\Sigma_{xx}]^{-1} \otimes \Pi_0 \,, \tag{2.29}$$

where Π_0 *is the (unique) solution of the* DLYAP *equation*

$$\Pi_0 = \tilde{A}_{yy} \Pi_0 \tilde{A}_{yy}^T + \tilde{\Sigma}_{yy|x} \,. \tag{2.30}$$

with

$$\tilde{\Sigma}_{yy|x} = \tilde{\Sigma}_{yy} - \tilde{\Sigma}_{yx}[\Sigma_{xx}]^{-1}\tilde{\Sigma}_{yx}^T = \begin{bmatrix} \Sigma_{yy|x} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix},$$

and $\Sigma_{yy|x}$ *the partial covariance matrix* (2.13).

Proof. See Supplementary Material, Section 2.7.3. □

We are now in a position to state our first principle result:

Theorem 1. *The asymptotic distribution of the single-regression Granger causality estimator under the null hypothesis $\theta \in \Theta_0$ is*

$$N\hat{F}^{SR}_{Y\to X} \xrightarrow{d} \chi^2\Big(I_{xx}\otimes\tilde{\Gamma}_{yy|x}, I_{xx}\otimes\big[\tilde{\Gamma}^{-1}\big]_{yy}\Big) \tag{2.31}$$

as $N\longrightarrow\infty$, where I_{xx} is the $n_x\times n_x$ identity matrix, and $\tilde{\Gamma}$, $\tilde{\Gamma}_{yy|x}$ satisfy the respective DLYAP *equations*

$$\tilde{\Gamma} = \tilde{A}\tilde{\Gamma}\tilde{A}^T + \tilde{\Sigma}\,, \qquad \tilde{\Gamma}_{yy|x} = \tilde{A}_{yy}\tilde{\Gamma}_{yy|x}\tilde{A}^T_{yy} + \tilde{\Sigma}_{yy|x}\,. \tag{2.32}$$

Proof. Eq. (2.30), *cf.* (2.4), specifies the autocovariance matrix (2.5) for a notional n_y-dimensional VAR(p) model with parameters $(A_{yy}, \Sigma_{yy|x})$. Accordingly, we write Π_0 as $\tilde{\Gamma}_{yy|x}$ from now on, so that (2.29) becomes $W_0(\theta) = 2[\Sigma_{xx}]^{-1}\otimes\tilde{\Gamma}_{yy|x}$, and from (2.24) and (2.26) it follows that

$$N\hat{F}^{\mathsf{SR}}_{Y\to X} \xrightarrow{d} \chi^2\Big([\Sigma_{xx}]^{-1}\otimes\tilde{\Gamma}_{yy|x}, \Sigma_{xx}\otimes\big[\tilde{\Gamma}^{-1}\big]_{yy}\Big)$$

as $N\longrightarrow\infty$. But from the transformation invariance $\chi^2(CAC^\mathsf{T}, B) = \chi^2(A, C^\mathsf{T}BC)$ and the mixed-product property of the Kronecker product, we may verify that the Σ_{xx} terms cancel (Σ_{xx} is positive-definite, and thus has an invertible Cholesky decomposition), and (2.31) follows. □

From (2.31) we see that the limiting distribution of $N\hat{F}^{\mathsf{SR}}_{Y\to X}$ is the sum of n_x random variables independently and identically distributed as $\chi^2\Big(\tilde{\Gamma}_{yy|x}, \big[\tilde{\Gamma}^{-1}\big]_{yy}\Big)$. By Supplementary Material, eq. 2.40, this distribution may be expressed in terms of the eigenvalues of $I_{xx}\otimes\Big(\big[\tilde{\Gamma}^{-1}\big]_{yy}\tilde{\Gamma}_{yy|x}\Big)$; these are the eigenvalues $\lambda_1,\ldots,\lambda_{pn_y}$ of $\big[\tilde{\Gamma}^{-1}\big]_{yy}\tilde{\Gamma}_{yy|x}$, where each λ_i appears with multiplicity n_x. The asymptotic distribution of $N\hat{F}^{\mathsf{SR}}_{Y\to X}$ under the null thus takes the form of a weighted sum of pn_y iid $\chi^2(n_x)$ variables:

$$\lambda_1 W_1 + \ldots + \lambda_{pn_y}W_{pn_y}\,, \qquad W_i \text{ iid} \sim \chi^2(n_x)\,.$$

The asymptotic mean and variance of $N\hat{F}^{\mathsf{SR}}_{Y\to X}$ for $\theta\in\Theta_0$ are

$$\mathrm{E}\Big[N\hat{F}^{\mathsf{SR}}_{Y\to X}\Big] \longrightarrow n_x\sum_{i=1}^{pn_y}\lambda_i\,, \qquad \mathrm{var}\Big[N\hat{F}^{\mathsf{SR}}_{Y\to X}\Big] \longrightarrow 2n_x\sum_{i=1}^{pn_y}\lambda_i^2 \tag{2.33}$$

respectively as $N\longrightarrow\infty$, from which the Γ-approximation of the generalized χ^2 distribution may be obtained [see Supplementary Material, eqs. 2.41 and 2.42]. Figure 2.1 plots generalized χ^2, Γ-approximation and empirical single-regression

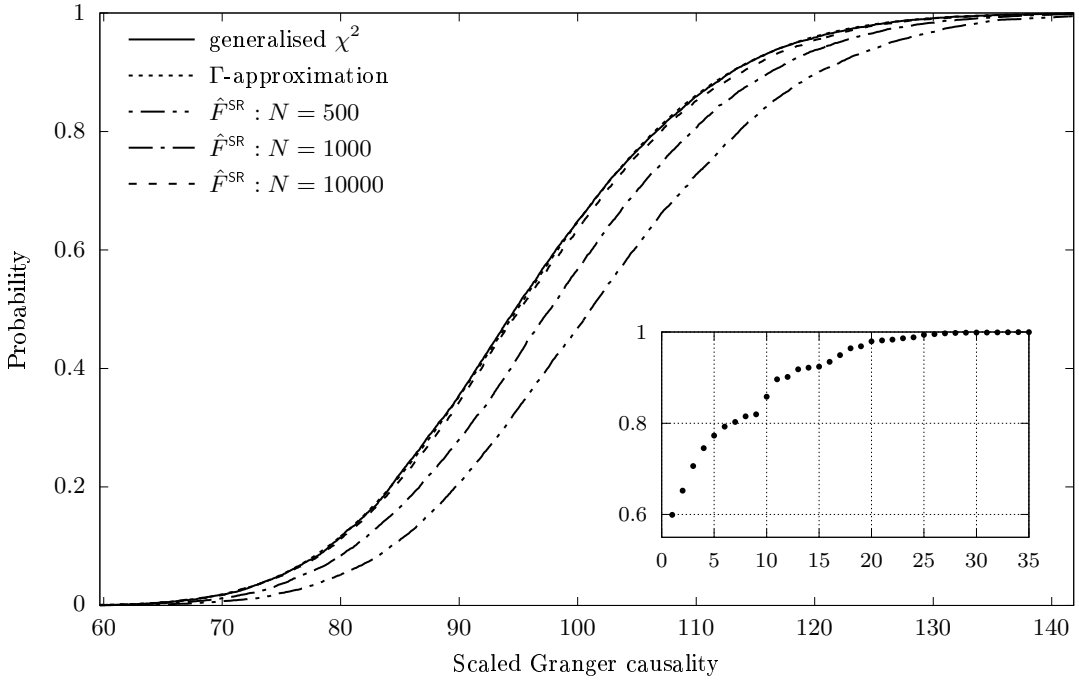

Fig. 2.1: Cumulative distributions for empirical single-regression Granger causality estimates and analytical distributions, for a representative null VAR model with $n_x = 3$, $n_y = 5$ and $p = 7$. Generalized χ^2 (solid line), Γ-approximation (dotted line, nearly indistinguishable from generalized χ^2), Granger causality estimator: $N = 10,000$ (dashes), $N = 1000$ (dot-dash), $N = 500$ (dot-dot-dash). The null VAR model was randomly generated according to the scheme described in Supplementary Material, Section 2.7.8, with spectral radius $\rho = 0.9$ and residuals generalized correlation $\gamma = 1$. Estimator plots are based on 10^4 generated time series. Inset figure: the $pn_y = 35$ distinct eigenvalues for the generalized χ^2 distribution, sorted by size. (Each eigenvalue will be repeated $n_x = 3$ times.)

estimator cumulative density functions (CDFs) for a representative null VAR model with $n_x = 3$, $n_y = 5$ and $p = 7$ for several sample sizes, illustrating asymptotic convergence with increasing sample length N. The Γ-approximation is barely distinguishable from the generalized χ^2. The eigenvalues λ_i are all > 0, since both $\left[\tilde{\Gamma}^{-1}\right]_{yy}$ and $\tilde{\Gamma}_{yy|x}$ are positive-definite. From Supplementary Material, eq. 2.42, we find that the shape parameter of the Γ-approximation satisfies $n_x/2 \leqslant \alpha \leqslant (pn_xn_y)/2$ and $\alpha = (pn_xn_y)/2 \iff$ all the λ_i are equal $\iff p = n_y = 1$, in which case the distribution of $N\hat{F}^{\text{SR}}_{Y\to X}$ is asymptotically $\chi^2(n_x)$ scaled by λ (*cf.* Supplementary Material, Section 2.7.7). We also state the following conjecture, which we have tested extensively empirically, but have so far been unable to prove rigorously:

Conjecture 1. *The eigenvalues of* $\left[\tilde{\Gamma}^{-1}\right]_{yy}\tilde{\Gamma}_{yy|x}$ *satisfy* $\lambda_i \leqslant 1$ $(i = 1, \ldots, pn_y)$.

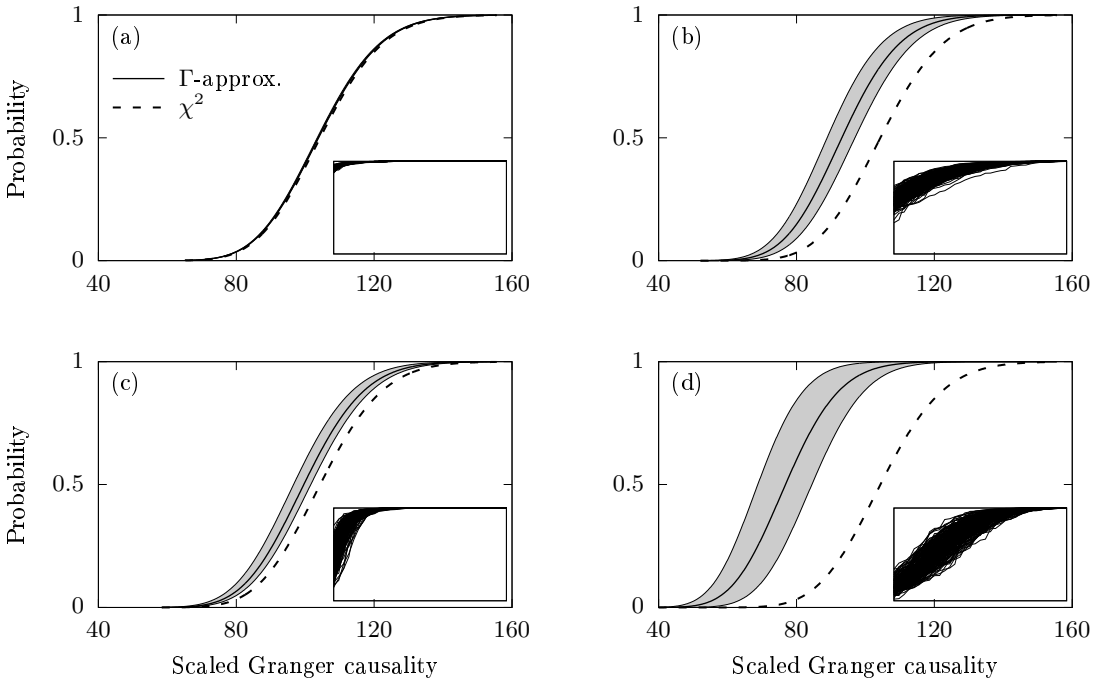

Fig. 2.2: Distribution of Γ-approximation cumulative distribution functions for a random sample of 200 null VAR models with $n_x = 3$, $n_y = 5$ and $p = 7$, for a selection of spectral radii ρ and residuals generalised correlation γ (see Supplementary Material, Section 2.7.8 for sampling details). At each scaled Granger causality value, solid lines plot the mean of the Γ-approximations, while shaded areas bound upper/lower 95% quantiles. Dashed lines plot the corresponding likelihood-ratio $\chi^2(d)$ distributions, with $d = pn_xn_y = 105$. (a) $\rho = 0.6, \gamma = 1$; (b) $\rho = 0.9, \gamma = 1$; (c) $\rho = 0.6, \gamma = 8$; (a) $\rho = 0.9, \gamma = 8$. Inset figures: the $pn_y = 35$ distinct eigenvalues sorted by size, for each of the 200 generalised χ^2 distributions (x-range is 1–35, y-range is 0–1).

If Conjecture 1 holds, then from Supplementary Material, eq. 2.42 and (2.33) the scale parameter of the Γ-approximation satisfies $0 \leqslant \beta \leqslant 2$. Simulations reveal that spectral radius and residuals generalised correlation of null parameters θ have a strong effect on the distribution of the eigenvalues λ_i. Spectral radius close to 1 and strong residuals cross-correlation give rise to a larger spread of eigenvalues < 1, resulting in asymptotic null sampling distributions significantly different from a (non-generalized) χ^2. Figure 2.2 presents the distribution of single-regression estimator CDFs under random sampling of VAR models of given size, spectral radius and residuals generalised correlation (see Supplementary Material, Section 2.7.8 for the VAR sampling scheme), where the effects of spectral radius and residuals correlation on the null distribution via the eigenvalues (inset figures) is clearly seen.

Assuming Conjecture 1, an immediate consequence of (2.33) and $N\hat{F}^{\mathsf{LR}}_{Y\to X} \xrightarrow{d} \chi^2(pn_xn_y)$ is that for $\theta \in \Theta_0$

$$\mathrm{E}\Big[N\hat{F}^{\mathsf{SR}}_{Y\to X}\Big] \leqslant \mathrm{E}\Big[N\hat{F}^{\mathsf{LR}}_{Y\to X}\Big], \qquad \mathrm{var}\Big[N\hat{F}^{\mathsf{SR}}_{Y\to X}\Big] \leqslant \mathrm{var}\Big[N\hat{F}^{\mathsf{LR}}_{Y\to X}\Big], \tag{2.34}$$

in the limit $N \to \infty$. As is apparent in Figure 2.2, the reduction in bias (rightward displacement of CDFs) and variance (slope of CDFs) is strongest for spectral radius close to 1 and for high residuals correlation.

2.4.3 The band-limited spectral single-regression estimator

We now consider the asymptotic null-distribution of the band-limited spectral Granger Causality estimator $\hat{f}_{Y\to X}(\mathcal{F})$ (2.14). It turns out that the null-hypothesis of vanishing $f_{Y\to X}(\mathcal{F})$ is in fact identical to the time-domain null-hypothesis (2.17). This can be seen by considering the point-frequency spectral Granger causality $f_{Y\to X}(\omega;\theta)$. Firstly, it is non-negative and clearly vanishes under H_0 for any ω. Further, by assumed stability of the VAR(p) (2.16), the inverse transfer function $\Phi(\omega)$ does not vanish anywhere, so that $\Psi(\omega)$, and hence, via (2.8), $S(\omega)$ and consequently $f_{Y\to X}(\omega;\theta)$, are analytic functions of the angular frequency ω [as well as of the $\theta = (A,\Sigma)$]. For a frequency range $\mathcal{F} \subseteq [0,2\pi]$ with measure $|\mathcal{F}| > 0$, then, $f_{Y\to X}(\mathcal{F};\theta)$ vanishes iff $f_{Y\to X}(\omega;\theta)$ is identically zero; i.e., precisely under the original null hypothesis H_0. This being the case, given a frequency range $\mathcal{F}$ we apply Proposition 1 to the asymptotic distribution of $\hat{f}_{Y\to X}(\mathcal{F})$ under the null hypothesis (2.17) $H_0 : A_{k,xy} = 0$ $(k = 1,\dots,p)$.

In the previous section we calculated the covariance $\Omega_0(\theta) = \Sigma_{xx} \otimes \big[\tilde{\Gamma}^{-1}\big]_{yy}$ of null parameters under H_0; it remains to calculate the null Hessian $W_0(\mathcal{F};\theta)$ for $f_{Y\to X}(\mathcal{F};\theta)$. Since (Lebesgue) integration and partial differentiation are linear operations, it follows that the Hessian on the original null space Θ_0 is just

$$W_0(\mathcal{F};\theta) = \frac{1}{|\mathcal{F}|}\int_{\mathcal{F}} W_0(\omega;\theta)\,d\omega\,,$$

where, for given ω, $W_0(\omega;\theta)$ is the Hessian of $f_{Y\to X}(\omega;\theta)$ on Θ_0 with respect to the null parameters $A_{k,xy}$ $(k = 1,\dots,p)$.

Proposition 4.

$$W_0(\mathcal{F};\theta) = [\Sigma_{xx}]^{-1} \otimes \mathfrak{Re}\big\{\tilde{S}_{yy|x}(\mathcal{F})\big\},$$

where

$$\tilde{S}_{yy|x}(\mathcal{F}) = \frac{1}{|\mathcal{F}|}\int_{\mathcal{F}} Z(\omega) \otimes S_{yy|x}(\omega)\, d\omega\,, \quad Z_{kk'}(\omega) = e^{-i\omega(k-k')}\ (k,k' = 1,\ldots,p) \tag{2.35}$$

with $S_{yy|x}(\omega)$ *the* CPSD *for a* VAR*(p) model with parameters* $(A_{yy}, \Sigma_{yy|x})$ *(cf. Theorem 1).*

Proof. See Supplementary Material, Section 2.7.5. □

We thus obtain our second principal result:

Theorem 2. *The asymptotic distribution of the single-regression band-limited Granger causality estimator over a frequency range* $\mathcal{F} \subseteq [0, 2\pi]$ *for* $\theta \in \Theta_0$ *is*

$$N\hat{f}_{Y\to X}(\mathcal{F}) \xrightarrow{d} \chi^2\Big(I_{xx} \otimes \mathfrak{Re}\big\{\tilde{S}_{yy|x}(\mathcal{F})\big\}, I_{xx} \otimes \big[\tilde{\Gamma}^{-1}\big]_{yy}\Big) \tag{2.36}$$

as $N \longrightarrow \infty$*, with* $\tilde{S}_{yy|x}(\mathcal{F})$ *as in* (2.35)*, and* $\tilde{\Gamma}$ *as in Theorem 1.*

Proof. The proof proceeds from Proposition 1 and Proposition 4 in the same way as the proof of Theorem 1. □

The matrix $\tilde{S}_{yy|x}(\mathcal{F})$ may be thought of as the spectral counterpart of the autocovariance matrix $\tilde{\Gamma}_{yy|x}$ of Section 2.4.2. In particular, for $\mathcal{F} = [0, 2\pi]$ we may confirm that $\tilde{S}_{yy|x}([0, 2\pi]) = \tilde{\Gamma}_{yy|x}$, so that that the distribution of $\hat{f}_{Y\to X}(\mathcal{F})$ is consistent with Theorem 1 and (2.15) for $\mathcal{F} = [0, 2\pi]$. The limiting asymptotic distribution of $\hat{f}_{Y\to X}([\omega - \varepsilon, \omega + \varepsilon])$ as $\varepsilon \to 0$ is obtained by simply replacing $\tilde{S}_{yy|x}(\mathcal{F})$ by $Z(\omega) \otimes S_{yy|x}(\omega)$ in (2.36); note that this is distinct from the distribution of the estimator $\hat{f}_{Y\to X}(\omega)$ under the point-frequency null $H_0(\omega)$ (see Section 2.6 for discussion of the point-frequency case). As before, the generalized χ^2 distribution in (2.36) may be described in terms of the eigenvalues λ_i $(i = 1,\ldots,pn_y)$ of $\big[\tilde{\Gamma}^{-1}\big]_{yy}\mathfrak{Re}\big\{\tilde{S}_{yy|x}(\mathcal{F})\big\}$. By Proposition 1 the λ_i are real and nonnegative, but in contrast to the time-domain case will not necessarily be asymptotically $\leqslant 1$ (*cf.* Conjecture 1). Empirically we observe that the maximum eigenvalue shrinks down towards 1 as the bandwidth $|\mathcal{F}|$ increases to 2π.

In Supplementary Material, Section 2.7.7 we present, as a worked example, a complete analysis of the single-regression Granger causality estimators in time and (band-limited) spectral domains, for the general bivariate VAR(1)

$$X_t = a_{xx}X_{t-1} + a_{xy}Y_{t-1} + \varepsilon_{xt} \tag{2.37a}$$
$$Y_t = a_{yx}X_{t-1} + a_{yy}Y_{t-1} + \varepsilon_{yt}\,. \tag{2.37b}$$

Model parameters are

$$A = \begin{bmatrix} a_{xx} & a_{xy} \\ a_{yx} & a_{yy} \end{bmatrix}, \qquad \Sigma = \mathrm{E}\big[\varepsilon_t \varepsilon_t^{\mathsf{T}}\big] = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix},$$

so that $\theta = (a_{xx}, a_{xy}, a_{yx}, a_{yy}, \sigma_{xx}, \sigma_{xy}, \sigma_{yy})$, and the null hypothesis H_0 is $a_{xy} = 0$.

2.5 Statistical inference with the single-regression estimators

2.5.1 Neyman-Pearson tests based on single regression estimators

Assuming a VAR(p) model, we construct a Neyman-Pearson test of the null hypothesis of zero Granger causality $H_0 : F_{Y\to X} = 0$ against the alternative of non-zero Granger causality $H_A : F_{Y\to X} \neq 0$. Since the proposed testing procedure is structurally identical no matter whether the time-domain statistic or the band-limited statistic is used (as noted in Section 2.4.3, the null hypothesis is the same in both cases), we adopt the following short-hand notation for brevity: F will refer to either the time-domain population Granger causality $F_{X\to Y}$ (2.21) or the band-limited Granger causality $f_{X\to Y}(\mathcal{F})$ (2.14) and, accordingly, $\hat{F}$ will refer to any single-regression Granger causality estimator.

A key difficulty in the construction is that the asymptotic distributions of the single-regression estimators $\hat{F}$ under the null hypothesis depend explicitly on the true null parameter. Our solution is as follows: having estimated the maximum-likelihood parameter $\hat{\theta} = (\hat{A}, \hat{\Sigma})$ for given time-series data, we "project" $\hat{\theta}$ onto the null space Θ_0, by setting $\hat{A}_{1,xy} = \ldots = \hat{A}_{p,xy} = 0$. Given a (not necessarily null) parameter θ, let Φ_θ denote the asymptotic CDF of the single-regression estimator—i.e., the generalized χ^2 of (2.31) (time domain) or (2.36) (band-limited)—evaluated at the projected parameter. Our test proceeds in three steps:

1. Calculate the maximum-likelihood VAR(p) parameter estimate $\hat{\theta}$.
2. Calculate the single-regression Granger causality estimate $\hat{F}$ based on $\hat{\theta}$.
3. Reject the null hypothesis if $\Phi_{\hat{\theta}}\big(N\hat{F}\big) > 1 - \alpha$.

Given $\theta \in \Theta_0$, the probability of a Type I error for the above test is

$$P_I(\theta; \alpha) = \mathrm{pr}\Big[\Phi_{\hat{\theta}}\left(N\hat{F}\right) > 1 - \alpha\Big]. \tag{2.38}$$

Note that both $\hat{\theta}$ and $\hat{F}$ are sample-size dependent random variables. The following result states that the proposed testing procedure is asymptotically valid:

Theorem 3. $\lim_{N\to\infty} P_I(\theta; \alpha) = \alpha$.

Proof. See Supplementary Material Section 2.7.6. □

The rate of convergence of $P_I(\theta; \alpha)$ to α can be expected to depend on the true null parameter θ (but note that this is also true of the likelihood-ratio test statistic).

2.5.2 Simulation results - time domain

To test statistical inference with the (time-domain) single-regression and likelihood-ratio estimators, we used the general bivariate VAR(1) (2.37), for which both $F_{Y\to X}$ and the sampling distributions under the null of its single-regression and likelihood-ratio estimators may be calculated analytically in closed form (Supplementary Material Section 2.7.7).

To compare the Type I error rate between the single-regression and likelihood-ratio tests, we simulated (2.37) with $a_{xy} = 0$ over a range of parameter values and sequence lengths N (without loss of generality we took Σ to be a correlation matrix with correlation κ). Results reveal very little difference between the performance of the respective estimators. Except for short sequence lengths ($N < 100$), Type I error rates for both tests are close to the significance level α, in line with the analysis in Section 2.5.1.

As regards statistical power, the Type II error rate given a *non*-null parameter $\theta \in \Theta$, is

$$P_{II}(\theta; \alpha) = \mathrm{pr}\Big[\Phi_{\hat{\theta}}\big(N\hat{F}\big) \leqslant 1 - \alpha\Big].$$

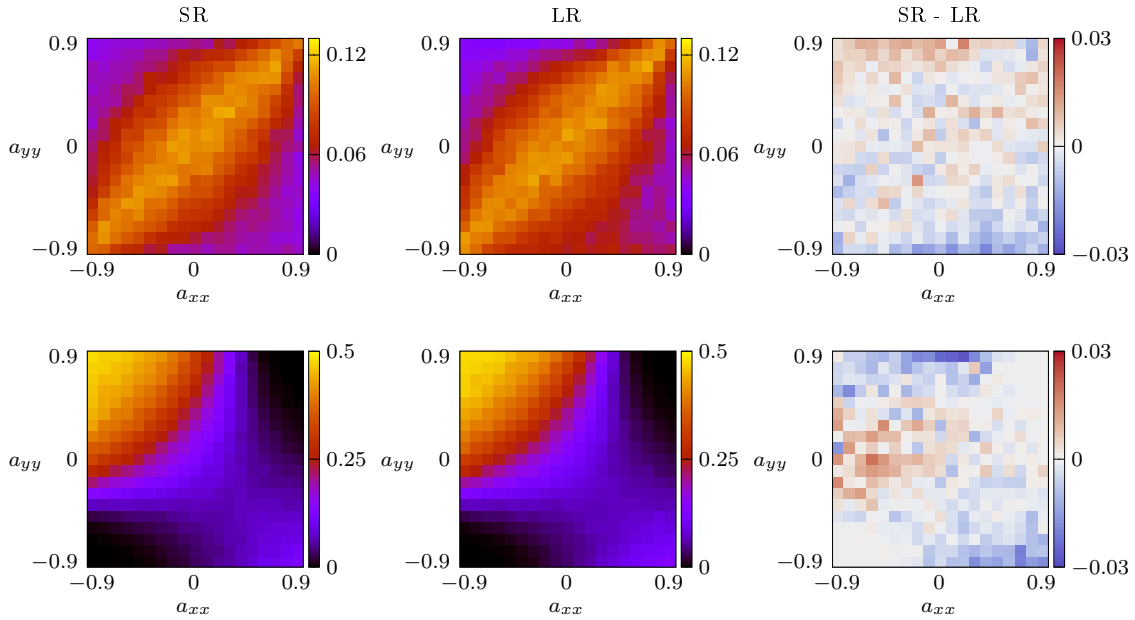

Fig. 2.3: Type II error rates (colour scale) at significance level $\alpha = 0.05$, based on $10,000$ realisations of the bivariate VAR(1) (2.37). Left column: single-regression test; centre column: likelihood-ratio test; right column: difference in error rate between estimators. Top row: $F = 0.01$, $a_{yx} = 0$, $\kappa = 0.5$, sequence length $N = 2^{10}$. Bottom row: $F = 0.001$, $a_{yx} = -1$, $\kappa = 0.5$, sequence length $N = 2^{12}$. In the right-column figures, red indicates higher statistical power for the likelihood-ratio test, while blue indicates higher statistical power for the single-regression test.

For the likelihood-ratio statistic we have the classical result due to Wald [44], which yields that the scaled estimator is approximately non-central $\chi^2(pn_xn_y; NF)$ in the large-sample limit, where F is the population Granger causality. The approximation only holds with reasonable accuracy for small values of F. In the single-regression case we have no equivalent result (but see discussion in Section 2.5.5). Clearly, $P_{II}(\theta; \alpha)$ will depend strongly on the population Granger causality value associated with specific parameters θ, but, as for the null case, will still vary within the subspace of parameters which yield a given population statistic; that is, for given $F > 0$, $P_{II}(\theta; \alpha)$ will vary over the set $\Theta_F = \{\theta \in \Theta : F(\theta) = F\}$.

To gain insight into comparative statistical power, we again simulated the bivariate VAR(1) (2.37) over a range of parameters, sequence lengths, and Granger causality values F. For given $F > 0$, a_{xx}, a_{yy} a_{yx} and κ, a_{xy} was calculated so that $F_{Y\to X} = F$ (see Supplementary Material Section 2.7.7). Figure 2.3 displays a comparison of Type II error rates between the single-regression and likelihood-ratio estimators for two illustrative model parameter regimes, (non-null) Granger causality values F, and data lengths N. We find that the difference in statistical power between the estimators is in general small, and that there are regions which favour one or the

other in roughly equal measure, the difference being greatest in regions of parameter space where spectral radius ρ is close to 1 and/or residuals correlation κ is large; in fact for any given region of parameter space, we found that it was generally possible to find a complementary region which "mirrored" the difference in statistical power between the estimators (e.g., by reversing the signs of a_{yx} and κ). It is not clear to what extent these conclusions extrapolate to higher model orders and system sizes, as increasing dimensionality of the parameter space renders detailed exploration impractical.

2.5.3 Utility of inference with the single-regression estimators

The above results suggest that as regards statistical inference, there is little to choose between the single-regression and likelihood-ratio tests, except in some "extreme" regions of parameter space where the difference in statistical power is sizeable. This raises the possibility of devising procedures for ascertaining, given empirical data, whether we are indeed in such a parameter regime. Given an estimated VAR model we might, for instance, use the model to generate surrogate time series, test the Type II error rate for the respective tests, and select the test with the smaller error rate. The extent to which this procedure might be confounded by model estimation error is, however, unclear; further research is required.

Regarding the band-limited estimator, even though the band-limited and time-domain null hypotheses are the same for any given frequency band $\mathcal{F}$, inference on $f_{Y\to X}(\mathcal{F})$ using the test statistic $N\hat{f}^{SR}_{X\to Y}(\mathcal{F})$ is nevertheless informative beyond a time-domain test based on $N\hat{F}^{SR}_{Y\to X}$, due to a difference in power profiles. Thus while the band-limited test may reject H_0 in the neighbourhood of ω_1 at some significance level, it may fail to reject H_0 at the same level around a different frequency ω_2, with the implication that while H_0 likely does not hold, there is likely to be a sizeable *contribution* to Granger causality around ω_1 while it is negligible around ω_2. By contrast, the time-domain test is insensitive to the localisation of Granger causality in the frequency spectrum and does therefore not allow any frequency-specific conclusions.

2.5.4 Unknown and infinite VAR model order

So far we have assumed that the model order of the underlying VAR model is both finite and known. However, these restrictions will generally not be met in practice. The question, then, is how statistical inference is affected when the true model order

is unknown, infinite or both. If the model order is unknown, statistical inference becomes a two-stage process: first we obtain a parsimonious model order estimate $\hat{p}$ using a standard selection procedure [61]. We then compute a test statistic $F^{(\hat{p})}$, the likelihood-ratio or single-regression estimate, using the selected model order; that is, maximum-likelihood VAR($\hat{p}$) parameter estimates are computed for the full (and in the case of the likelihood-ratio estimator, reduced) models.

The central question is how the model order selection step should be performed so that the two-step testing procedure as a whole is (1) asymptotically valid, and (2) as statistically powerful as possible. We consider first the implications of selecting a fixed model order $q \neq p$ for inference on a finite-order VAR(p) process. If $q < p$, then the asymptotic statements underlying the likelihood-ratio test (i.e. Wilks' theorem) and the single-regression test described in Section 2.5.1 no longer hold; thus, for instance, in the case of the likelihood-ratio statistic, $NF^{(q)} \xrightarrow{d} \chi^2(qn_x n_y)$ fails under the null hypothesis (2.17). On the other hand, if $q > p$ then Wilks' theorem *does* apply, because (*cf.* Section 2.3.2) one can always subsume a VAR model by a higher-order model. For fixed $q > p$ then, $NF^{(q)} \xrightarrow{d} \chi^2(qn_x n_y)$ under the null hypothesis even though the model is over-specified. However, simulation results suggest two problems: firstly, the rate of convergence of the test statistic decreases with q (potentially leading to inflated Type I errors in small samples), and secondly, statistical power is degraded.

Because the reduced process will in general be infinite order, the reduced model order estimate will diverge to infinity as sample size increases, leading to suboptimal inference as described above. This implies that for the purposes of statistical inference (as opposed to estimation of effect size), model order should be selected for the *full*, rather than reduced process. There remains a choice regarding which of the many possible selection criteria should be deployed. If the model order selection criterion utilised is *consistent* then the probability of choosing the correct model order converges to 1 as $N \to \infty$. It is not hard to show that in this case the two-step procedure consisting of model selection followed by a Neyman-Pearson test is asymptotically valid. We note that the popular Akaike Information Criterion (AIC) is *not* consistent, whereas, e.g., Schwartz' Bayesian Information Criterion (BIC) and Hannan and Quinn's Information Criterion [HQIC; 68] are consistent.

As regards the infinite-order VAR case, establishing an asymptotically-valid scheme seems more difficult, and merits further research. Preliminary experiments indicate that, at least for vector autoregressive moving-average (VARMA; equivalently state-space) processes, the single-regression estimator with consistent model order

selection yields asymptotically valid inference, in the sense that the Type I error rate converges to a specified significance level α.

Further research is required to explore more fully the consequences of model order selection on statistical inference. To this end, The VARMA representations in Dufour and Pelletier [69] may be particularly appropriate, since they are readily identifiable and (unlike innovations-form state-space models) easily specified in a form which makes causal interactions explicit.

2.5.5 The alternative hypothesis

We may consider two approaches to approximating the sampling distribution of the time-domain and band-limited spectral estimators, which address two distinct scenarios. In the first scenario, we suppose given a fixed true parameter $\theta \notin \Theta_0$, and consider the asymptotic sampling distribution of the Granger causality statistic as sample length $N \to \infty$. In this case, the preconditions of Proposition 1 do not apply; in particular, the gradient of the statistic will not in general vanish at θ, so that a 1st-order multivariate Delta Method is appropriate. This yields a normal distribution for the estimator, with mean equal to the population Granger causality. If the statistic is $f\left(\hat{\theta}\right)$, then explicitly we have

$$N^{\frac{1}{2}}\left[f\left(\hat{\theta}\right) - f(\theta)\right] \xrightarrow{d} \mathcal{N}\left(0, \nabla f(\theta) \cdot \Omega(\theta) \cdot \nabla f(\theta)^{\mathsf{T}}\right).$$

The variance $\sigma^2 = \nabla f(\theta) \cdot \Omega(\theta) \cdot \nabla f(\theta)^{\mathsf{T}}$ may be computed from the known form of the statistic, although the gradients are harder to calculate, since (i) in the time domain the DARE does not, as in the null case (Section 2.4.2) collapse to a DLYAP equation (2.30), while (ii) in the spectral band-limited case (Section 2.4.3), the transfer function $\Psi(\omega; \theta)$ is no longer block-triangular. Gradients, furthermore, must be calculated with respect to all (rather than just null) parameters. This scenario is more pertinent in a realistic empirical situation where, for instance, we are reasonably confident (via a Neyman-Pearson test as described in Section 2.5.1) that an estimated Granger causality is significantly different from zero, and we would like to put confidence bounds on the estimate.

Under the second and more difficult to analyse scenario, we suppose that sample length N is fixed (but large), and we consider the limiting distribution of the single-regression Granger causality estimator as the true non-null parameter approaches the null subspace Θ_0. We are now in the regime of Wald's Theorem [44], where the asymptotics of the Taylor expansion on which the 1st- and 2nd-order Delta

Methods are based become a balancing act between sample length N and the distance between the true parameter and the null subspace. This is likely to be difficult to calculate; we conjecture that (by analogy with Wald's Theorem) the asymptotic distribution will be a non-central generalized χ^2. This scenario is more pertinent to analysis of statistical power (*cf.* Section 2.5.2).

2.6 Extensions and future research directions

Any estimator of the form $g(\hat{\theta})$, where $\hat{\theta}$ is the maximum-likelihood parameter estimator, will converge in distribution to a generalized χ^2 distribution under the associated null hypothesis $\theta \in \Theta_0$ if the population statistic $g(\theta)$ satisfies the pre-requisites for Proposition 1. This covers a range of extensions to our results. They vary in tractability according to the difficulty of explicit calculation of the Fisher information matrix $\Omega(\theta)$ and Hessian $W(\theta)$ for $\theta \in \Theta_0$.

The conditional case: Extending the time-domain Theorem 1 to the conditional case [11] is reasonably straightforward. Given a partitioning $U_t = [X_t^\mathsf{T}\ Y_t^\mathsf{T}\ Z_t^\mathsf{T}]^\mathsf{T}$ of the variables, the time-domain conditional population Granger causality statistic is

$$F_{Y\to X|Z}(\theta) = \log\left|\Sigma^{\mathsf{R}}_{xx}\right| - \log\left|\Sigma_{xx}\right|,$$

where now the reduced autoregression (2.9) is on $[X_t^\mathsf{T}\ Z_t^\mathsf{T}]^\mathsf{T}$ rather than just X_t. Again, Σ^{R} is a deterministic (albeit more complicated) function $V(\theta)$ of the VAR parameters, which may again be expressed in terms of a DARE [18]. Although more complex, derivation of the appropriate Hessian proceeds along the same lines as in Section 2.4.2. Extension to the conditional case in the frequency domain (band-limited estimator) is more challenging, due to the complexity of the statistic; see e.g., Barnett and Seth [18, Sec. II]. While the unconditional spectral statistic only references the full model parameters, in the conditional case both full and reduced model parameters are required.

The spectral point-frequency estimator: The null hypothesis $H_0(\omega)$ for vanishing of $f_{Y\to X}(\omega;\theta)$ (2.12) at the point frequency ω is $\Psi_{xy}(\omega) = 0$, where $\Psi(\omega)$ is the transfer function (2.6) for the VAR model, or [10]:

$$H_0(\omega) : \sum_{k=1}^{p} A_{k,xy} e^{-ik\omega} = 0\,. \tag{2.39}$$

For given ω, (2.39) represents $2n_xn_y$ linear constraints on the pn_xn_y regression coefficient matrices A_k. Note that for $p \leqslant 2$, if $\omega \neq 0$, π, or 2π then $H_0(\omega)$ coincides with the original $H_0 : A_{k,xy} = 0$ (*cf.* Supplementary Material, Section 2.7.7). Calculation of the point-frequency asymptotic sampling distribution is in principle approachable via a similar technique as before (Proposition 1 must be adjusted for the case of more general linear constraints). However, we contend that in real-world applications it makes more sense in any case to consider inference on spectral Granger causality on a (possibly narrow-band) frequency range via the band-limited spectral Granger causality $f_{Y\to X}(\mathcal{F};\theta)$ (2.14) as discussed in Section 2.4.3 rather than at point frequencies. Firstly, for a given VAR(p), if the broadband null condition H_0 is not satisfied, then the point-frequency null condition $H_0(\omega)$ will only be satisfied precisely at most at a finite number of (in practice unknown) point frequencies. Secondly, real-world spectral phenomena are likely to be to some extent broadband [e.g., power spectra of neural processes [70]] and/or otherwise blurred by noise. The asymptotic sampling distribution of the point-frequency estimator is nonetheless at least of academic interest, since as far as the authors are aware, it remains unknown.

The state-space Granger causality statistic: The state-space Granger causality statistic, unconditional and conditional, in time and frequency domains, was introduced in Barnett and Seth [18] [see also 51]. It is an attractive alternative to VAR-based Granger causality, but its sampling distribution under the null hypothesis of vanishing Granger causality remains unknown. The state-space approach extends Granger causality estimation (and, potentially, inference) from the class of finite-order VAR processes to the super-class of state-space (equivalently finite-order VARMA) processes. The power of the method derives from the fact that (i) unlike the class of finite-order VAR models, the class of state-space models is closed under subprocess extraction (an essential ingredient of the Granger causality construct), and (ii) many real-world data, notably econometric and neurophysiological, have a strong moving-average component, and are thus more parsimoniously modelled as VARMA rather than pure VAR processes. The class of state-space models is in addition (again in contrast to the finite-order VAR class) closed under sub-sampling, temporal/spatial aggregation and additive observation noise – all common features of real-world data acquisition and observation procedures.

Solo [51] states without proof that the asymptotic distribution of the state-space estimator will be a simple χ^2 under the null hypothesis. However, like the VAR single-regression Granger causality statistic, the state-space statistic is also a non-negative deterministic function of the model parameters, so that the 2nd-order Delta Method (Proposition 1) applies, and the sampling distribution of the estimator

under the null will thus be a *generalised* χ^2. This explains the simulation-based observation in Barnett and Seth [18] that the state-space estimator under the null is well-approximated by a Γ distribution. In comparison with the VAR case, there are two challenges to calculation of the generalised χ^2 parameters: (i) calculation of the Fisher information matrix, and (ii) non-linearity of the null condition [18, eq. 17]. While (ii) complicates calculation of the Hessian (*cf.* Supplementary Material, Section 2.7.2), (i) is likely to present a more formidable obstacle, due to the considerable complexity of closed-form expressions for the Fisher information matrix [71].

2.7 Supplementary Materials

2.7.1 The generalized χ^2 family of distributions

Let $Z \sim \mathcal{N}(0, B)$ be a zero-mean n-dimensional multivariate-normal random vector with covariance matrix B, and A an $n \times n$ symmetric matrix. Then [72] we write $\chi^2(A, B)$ for the distribution of the random quadratic form $Q = Z^\mathsf{T} A Z$. If $A = B = I$, then $\chi^2(A, B)$ reduces to the usual $\chi^2(n)$. If A is $m \times m$ and C is $m \times n$, then $\chi^2(A, CBC^\mathsf{T}) = \chi^2(C^\mathsf{T} AC, B)$.

It is not hard to show [65] that if B is positive-definite and A symmetric (which will be the case for the generalized χ^2 distributions we encounter), then $\chi^2(A, B) = \chi^2(\Lambda, I)$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_1, \ldots, \lambda_n$ the eigenvalues of BA, or, equivalently, of RAR^T where R is the right Cholesky factor of B (so that $R^\mathsf{T} R = B$). In that case, we have

$$\lambda_1 U_1^2 + \ldots + \lambda_n U_n^2 \sim \chi^2(A, B)\,, \qquad U_i \text{ iid} \sim \mathcal{N}(0, 1)\,, \tag{2.40}$$

so that $\chi^2(A, B)$ is a weighted sum of independent χ^2-distributed variables, and in particular if the λ_i are all equal then we have a scaled $\chi^2(n)$ distribution. From (2.40), moments of a generalized χ^2 variable may be conveniently expressed in terms of the eigenvalues; thus we may calculate that for $Q \sim \chi^2(A, B)$

$$\mathrm{E}[Q] = \mu \;=\; \sum_{i=1}^{n} \lambda_i\,, \qquad \mathrm{var}[Q] = \sigma^2 = 2 \sum_{i=1}^{n} \lambda_i^2\,. \tag{2.41}$$

Empirically, it is found that generalized χ^2 variables (at least for A symmetric and B positive-definite) are very well approximated by Γ distributions: specifically, we have $Q \approx \Gamma(\alpha, \beta)$ with shape and scale parameters

$$\alpha = \mu^2 \sigma^{-2}\,, \qquad \beta = \mu^{-1}\sigma^2 \tag{2.42}$$

respectively.

2.7.2 Proof of Main Article, Proposition 1

Proof. Let $\theta = [x^\mathsf{T}\ y^\mathsf{T}]^\mathsf{T}$ where $x_i = \theta_i$ $(i = 1, \ldots, d)$ and $y_j = \theta_{d+j}$ $(j = 1, \ldots, s)$. Since by definition $f(0, y) = 0$ for all y, we have immediately $\nabla_y f(0, y) = 0$ for all y. Treating y as fixed, we expand $f(x, y)$ in a Taylor series around $x = 0$:

$$f(x, y) = \nabla_x f(0, y)x + \tfrac{1}{2}x^\mathsf{T}\nabla^2_{xx} f(0, y)x + \tfrac{1}{2}x^\mathsf{T}K(x, y)x\,, \tag{2.43}$$

where for fixed y, $K(x, y)$ is a $d \times d$ matrix function of x, and $\lim_{x\to 0} \|K(x, y)\| = 0$. Now we show that since $f(x, y)$ is non-negative, we must have $\nabla_x f(0, y) = 0$ for all y. Suppose, say, $\nabla_{x_1} f(0, y) = -g < 0$. Setting $x_1 = \varepsilon > 0$ [if $\nabla_{x_1} f(0, y) > 0$ we take $x_1 = -\varepsilon$] and $x_2 = \ldots = x_d = 0$, (2.43) yields

$$f(x, y) = -g\varepsilon + \tfrac{1}{2}\left[\nabla^2_{x_1x_1} f(0, y) + K_{11}(x, y)\right]\varepsilon^2.$$

Now since $\lim_{\varepsilon\to 0} \|K(x, y)\| = 0$, we can always choose ε small enough that $\frac{1}{2}\left[\nabla^2_{x_1x_1} f(0, y) + K_{11}(x, y)\right]\varepsilon < g$, so that $f(\varepsilon, 0, ..., 0, y) < 0$, a contradiction. Thus we have $\nabla_x f(0, y) = 0$ for all y, proving Proposition 1a.

From (2.43) we thus have

$$f(x, y) = \tfrac{1}{2}x^\mathsf{T}\nabla^2_{xx} f(0, y)x + \tfrac{1}{2}x^\mathsf{T}K(x, y)x\,. \tag{2.44}$$

To see that $\nabla^2_{xx} f(0, y)$ must be positive-semidefinite, we assume the contrary. We may then find a unit d-dimensional vector u such that $u^\mathsf{T}\nabla^2_{xx} f(0, y)u = -G < 0$. Setting $x = \varepsilon u$, we may then choose ε small enough that $u^\mathsf{T}K(\varepsilon u, y)u < G$, so that again $f(x, y)$ is negative and we have a contradiction. Finally, $\nabla^2_{xy} f(0, y) = \nabla^2_{yy} f(0, y) = 0$ for all y follows directly from (2.44), and we have established Proposition 1b.

We now prove Proposition 1c using a 2nd-order Delta Method [50]. Let $\theta \in \Theta_0$. Since $f(\theta)$ and its gradient $\nabla f(\theta)$ both vanish, the Taylor expansion of $f(\vartheta_N)$ around θ takes the form

$$f(\vartheta_N) = \tfrac{1}{2}\,(\vartheta_N - \theta)^{\mathsf{T}}\, W(\theta)\,(\vartheta_N - \theta) + (\vartheta_N - \theta)^{\mathsf{T}}\, K(\vartheta_N)\,(\vartheta_N - \theta)\,, \tag{2.45}$$

where $W(\theta) = \nabla^2 f(\theta)$ is the Hessian matrix of f evaluated at θ, and $\lim_{\theta' \to \theta} \|K(\theta')\| = 0$. By assumption $N^{\frac{1}{2}}(\vartheta_N - \theta) \xrightarrow{d} Z$ as $N \to \infty$, where $Z \sim \mathcal{N}(0, \Omega(\theta))$. Therefore, multiplying both sides of (2.45) by the sample size N, by the Continuous Mapping Theorem [73], we have

$$N f(\vartheta_N) \xrightarrow{d} \tfrac{1}{2} Z^{\mathsf{T}} W(\theta) Z \tag{2.46}$$

as $N \longrightarrow \infty$, and Proposition 1c follows immediately from (2.46) and Proposition 1b. □

In the more general case where the null manifold Θ_0 is a smooth s-dimensional submanifold of $\Theta \subseteq \mathbb{R}^r$, we can always find, at least locally, a change of coordinates $\psi : \mathbb{R}^r \to \mathbb{R}^r$ such that in the new coordinate system $\tilde{\Theta}_0 = \psi(\Theta_0)$ and $\tilde{f} = f \circ \psi^{-1}$ satisfy the criteria of Proposition 1. It is not hard to show then that (2.45) holds for $W(\theta) = \nabla\psi(\theta)^{\mathsf{T}} \cdot \tilde{W}(\theta) \cdot \nabla\psi(\theta)$ [note that the Jacobian matrix $\nabla\psi(\theta)$ is invertible] and $\tilde{W}(\theta)$ takes the block form of Proposition 1b. We thus obtain

$$N f(\vartheta_N) \xrightarrow{d} \chi^2\big(\tfrac{1}{2}\tilde{W}_0(\theta), \tilde{\Omega}_0(\theta)\big) \tag{2.47}$$

as $N \longrightarrow \infty$, where $\tilde{\Omega}_0(\theta) = \nabla\psi(\theta)^{\mathsf{T}} \cdot \Omega(\theta) \cdot \nabla\psi(\theta)$.

2.7.3 Proof of Main Article, Proposition 2

Proof. Following Barnett and Seth [18], given a VAR(p) model

$$U_t = \sum_{k=1}^{p} A_k U_{t-k} + \varepsilon_t \tag{2.48}$$

with parameters $\theta = (A; \Sigma)$, we create an equivalent innovations-form state-space model [56]

$$\begin{aligned} Z_{t+1} &= \tilde{A} Z_t + K\varepsilon_t\,, \\ U_t &= A Z_t \;+\; \varepsilon_t\,, \end{aligned}$$

where

$$\tilde{A} = \begin{bmatrix} A_1 & A_2 & \dots & A_{p-1} & A_p \\ I & 0 & \dots & 0 & 0 \\ 0 & I & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & I & 0 \end{bmatrix}, \quad A = \begin{bmatrix} A_1 & A_2 & \dots & A_{p-1} & A_p \end{bmatrix}, \quad K = \begin{bmatrix} I \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

$\tilde{A}$ is the $pn \times pn$ state transition matrix (the companion matrix of VAR cofficients) for the VAR(p) (2.48)], K the $pn \times n$ Kalman gain matrix, and A the $n \times pn$ observation matrix. As before, we use subscript x for the indices $1, \dots, n_x$, y for the indices $n_x + 1, \dots, n$, and we use an asterisk to denote "all indices". The subprocess X_t then satisfies the state-space model

$$\begin{aligned} Z_{t+1} &= \tilde{A} Z_t \quad + K\varepsilon_t \,, \\ X_t &= A_{x*} Z_t + \quad \varepsilon_{x,t} \,. \end{aligned}$$

This state-space model is no longer in innovations form; we can, however [see 18] derive an innovations-form state-space model for X_t by solving the discrete-time algebraic Riccati equation (DARE)

$$P = \tilde{A} P \tilde{A}^\mathsf{T} + \tilde{\Sigma} - \left(\tilde{A} P A_{x*}^\mathsf{T} + \tilde{\Sigma}_{*x}\right)\left(A_{x*} P A_{x*}^\mathsf{T} + \Sigma_{xx}\right)^{-1}\left(\tilde{A} P A_{x*}^\mathsf{T} + \tilde{\Sigma}_{*x}\right)^\mathsf{T} \quad (2.51)$$

for P (a $pn \times pn$ symmetric matrix), with

$$\tilde{\Sigma} = \begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}, \qquad \tilde{\Sigma}_{*x} = \begin{bmatrix} \Sigma_{*x} \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

which are, respectively, $pn \times pn$ and $pn \times n_x$. We note that under our assumptions, a stabilising solution for (2.51) exists, and is unique [51]. Then

$$\begin{aligned} Z_{t+1} &= \tilde{A} Z_t \quad + K^\mathsf{R} \varepsilon_t^\mathsf{R} \,, && (2.52a) \\ X_t &= A_{x*} Z_t + \quad \varepsilon_t^\mathsf{R} && (2.52b) \end{aligned}$$

is in innovations form, with innovations covariance matrix and Kalman gain matrix

$$\begin{aligned} \Sigma^\mathsf{R} &= A_{x*} P A_{x*}^\mathsf{T} + \Sigma_{xx} \,, && (2.53a) \\ K^\mathsf{R} &= \left(\tilde{A} P A_{x*}^\mathsf{T} + \tilde{\Sigma}_{*x}\right)\left[\Sigma^\mathsf{R}\right]^{-1} && (2.53b) \end{aligned}$$

respectively. The innovations $\varepsilon_t^{\mathsf{R}}$ in (2.52) are precisely the residuals of the reduced VAR model for X_t, and $\Sigma^{\mathsf{R}} = \mathrm{E}[\varepsilon_t^{\mathsf{R}}\varepsilon_t^{\mathsf{RT}}]$ implicitly defines $V(\theta)$ as required for the single-regression Granger causality statistic $F^{\mathsf{SR}}_{Y\to X}(\theta)$ (see Main Article, Section 2.3 and Section 2.3.3).

We may in fact confirm that

$$\Sigma^{\mathsf{R}} = V(\theta) = A_{xy}\Pi A_{xy}^{\mathsf{T}} + \Sigma_{xx}\,, \tag{2.54}$$

where the $pn_y \times pn_y$ symmetric matrix Π is the unique stabilising solution of the "reduced DARE"

$$\Pi = \tilde{A}_{yy}\Pi\tilde{A}_{yy}^{\mathsf{T}} + \tilde{\Sigma}_{yy} - \big(\tilde{A}_{yy}\Pi A_{xy}^{\mathsf{T}} + \tilde{\Sigma}_{yx}\big)\big(A_{xy}\Pi A_{xy}^{\mathsf{T}} + \Sigma_{xx}\big)^{-1}\big(\tilde{A}_{yy}\Pi A_{xy}^{\mathsf{T}} + \tilde{\Sigma}_{yx}\big)^{\mathsf{T}}, \tag{2.55}$$

with

$$\tilde{A}_{yy} = \begin{bmatrix} A_{1,yy} & A_{2,yy} & \dots & A_{p-1,yy} & A_{p,yy} \\ I & 0 & \dots & 0 & 0 \\ 0 & I & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & I & 0 \end{bmatrix}, \quad A_{xy} = \begin{bmatrix} A_{1,xy} & A_{2,xy} & \dots & A_{p-1,xy} & A_{p,xy} \end{bmatrix},$$

which are, respectively, $pn_y \times pn_y$ and $n_x \times pn_y$, and

$$\tilde{\Sigma}_{yy} = \begin{bmatrix} \Sigma_{yy} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}, \qquad \tilde{\Sigma}_{yx} = \begin{bmatrix} \Sigma_{yx} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

respectively, $pn_y \times pn_y$ and $pn_y \times n_x$. To see this, we may verify by substitution that if

$$\Pi = \begin{bmatrix} \Pi_{11} & \cdots & \Pi_{1p} \\ \vdots & & \vdots \\ \Pi_{p1} & \cdots & \Pi_{pp} \end{bmatrix}$$

solves the reduced-dimension DARE (2.55), where the Π_{kl} are $n_y \times n_y$, then

$$P = \begin{bmatrix} \begin{bmatrix} 0_{n_x\times n_x} & 0_{n_x\times n_y} \\ 0_{n_y\times n_x} & \Pi_{11} \end{bmatrix} & \cdots & \begin{bmatrix} 0_{n_x\times n_x} & 0_{n_x\times n_y} \\ 0_{n_y\times n_x} & \Pi_{1p} \end{bmatrix} \\ \vdots & & \vdots \\ \begin{bmatrix} 0_{n_x\times n_x} & 0_{n_x\times n_y} \\ 0_{n_y\times n_x} & \Pi_{p1} \end{bmatrix} & \cdots & \begin{bmatrix} 0_{n_x\times n_x} & 0_{n_x\times n_y} \\ 0_{n_y\times n_x} & \Pi_{pp} \end{bmatrix} \end{bmatrix}$$

solves the original DARE (2.51), and that Σ^{R} is given by (2.54). We may also confirm that the Kalman gain matrix (2.53b) for the reduced DARE is

$$K^{\mathsf{R}} = \begin{bmatrix} I_{n_x \times n_x} \\ L_1^{\mathsf{R}} \\ 0_{n_x \times n_x} \\ L_2^{\mathsf{R}} \\ \vdots \\ 0_{n_x \times n_x} \\ L_p^{\mathsf{R}} \end{bmatrix},$$

where

$$L^{\mathsf{R}} = \left(\tilde{A}_{yy} \Pi A_{xy}^{\mathsf{T}} + \tilde{\Sigma}_{yx}\right) \left[\Sigma^{\mathsf{R}}\right]^{-1} = \begin{bmatrix} L_1^{\mathsf{R}} \\ L_2^{\mathsf{R}} \\ \vdots \\ L_p^{\mathsf{R}} \end{bmatrix}$$

(the L_k^{R} are $n_y \times n_x$) is the Kalman gain matrix associated with the DARE (2.55). □

2.7.4 Proof of Main Article, Theorem 1

Proof. To calculate $W_0(\theta)$ we require derivatives up to 2nd order of $V(\theta)$ (2.54) with respect to the null-hypothesis parameters (that is, with respect to $A_{k,ij}$ for $i \in x$, $j \in y$), evaluated for $\theta \in \Theta_0$. From (2.54) we may calculate:

$$\frac{\partial V_{ii'}}{\partial A_{k,uv}} = \delta_{ui} \left[\Pi A_{xy}^{\mathsf{T}}\right]_{k,vi'} + \delta_{ui'} \left[A_{xy} \Pi\right]_{k,iv} + \left[A_{xy} \frac{\partial \Pi}{\partial A_{k,uv}} A_{xy}^{\mathsf{T}}\right]_{ii'}, \tag{2.56}$$

where indices $i, i', u, u' \in x$, indices $j, j', v, v' \in y$ and $k, k' = 1, \ldots, p$. Since A_{xy} vanishes under the null hypothesis, we have

$$\left.\frac{\partial V_{ii'}}{\partial A_{k,uv}}\right|_{\theta \in \Theta_0} = 0\,, \tag{2.57}$$

and from (2.56) we find

$$\left.\frac{\partial^2 V_{ii'}}{\partial A_{k,uv} \partial A_{k',u'v'}}\right|_{\theta \in \Theta_0} = \left[\delta_{ui}\delta_{u'i'} + \delta_{ui'}\delta_{u'i}\right] \Pi_{kk',vv'}. \tag{2.58}$$

We see then that Π is required only on the null space $A_{xy} = 0$, in which case the DARE (2.55) becomes a discrete Lyapunov (DLYAP) equation for $\Pi_0 = \Pi|_{\theta\in\Theta_0}$:

$$\Pi_0 = \tilde{A}_{yy}\Pi_0\tilde{A}_{yy}^\intercal + \tilde{\Sigma}_{yy} - \tilde{\Sigma}_{yx}[\Sigma_{xx}]^{-1}\tilde{\Sigma}_{yx}^\intercal.$$

We may now calculate the required Hessian. For null parameters $\theta_\alpha, \theta_\beta$, from the definition $F_{Y\to X}(\theta) = \log|V(\theta)| - \log|\Sigma_{xx}|$ and using (2.57) we may calculate

$$\left.\frac{\partial^2 F_{Y\to X}}{\partial\theta_\alpha\partial\theta_\beta}\right|_{\theta\in\Theta_0} = \left.\frac{\partial^2 \log|V|}{\partial\theta_\alpha\partial\theta_\beta}\right|_{\theta\in\Theta_0} = \mathrm{tr}\left[[\Sigma_{xx}]^{-1}\left.\frac{\partial^2 V}{\partial\theta_\alpha\partial\theta_\beta}\right|_{\theta\in\Theta_0}\right]. \tag{2.59}$$

Eq. (2.58) then yields

$$[W_0(\theta)]_{[k,uv],[k',u'v']} = 2\left[[\Sigma_{xx}]^{-1}\right]_{uu'}[\Pi_0]_{kk',vv'}\,,$$

or

$$W_0(\theta) = 2[\Sigma_{xx}]^{-1}\otimes\Pi_0\,,$$

as required. □

2.7.5 Proof of Main Article, Proposition 4

Proof. Dropping the "ω" and "θ" arguments for compactness where convenient, on the null space Θ_0 we have $\Phi_{xy} = 0$, and since then Φ is lower block-triangular, we have also $\Psi_{xx} = [\Phi_{xx}]^{-1}$, $\Psi_{yy} = [\Phi_{yy}]^{-1}$, and $\Psi_{xy} = 0$. The CPSD for the process X_t is given by

$$S_{xx} = [\Psi S\Psi^*]_{xx} = \Psi_{xx}\Sigma_{xx}\Psi_{xx}^* + \Psi_{xy}\Sigma_{yx}\Psi_{xx}^* + \Psi_{xx}\Sigma_{xy}\Psi_{xy}^* + \Psi_{xy}\Sigma_{yy}\Psi_{xy}^*\,.$$

On the null space $S_{xx} = \Psi_{xx}\Sigma_{xx}\Psi_{xx}^*$ so that $[S_{xx}]^{-1} = \Phi_{xx}^*[\Sigma_{xx}]^{-1}\Phi_{xx}$,

We define $T(\omega)$ to be the $n_x \times n_x$ (Hermitian) matrix $T(\omega) = \Psi_{xy}(\omega)\Sigma_{yy|x}\Psi_{xy}(\omega)^*$, so that from Main Article, eq. 2.12, $f_{Y\to X}(\omega) = \log|S_{xx}(\omega)| - \log|S_{xx}(\omega) - T(\omega)|$. $T(\omega)$ vanishes on the null space. We may check that

$$\frac{\partial\Psi_{pq}}{\partial A_{k,rs}} = \Psi_{pr}\Psi_{sq}e^{-i\omega k} \qquad (p,q,r,s = 1,\ldots,n;\ k = 1,\ldots,p)\,, \tag{2.60}$$

from which we may calculate (with $i, i', u, u' \in x$, $j, j', v, v' \in y$)

$$\frac{\partial T_{ii'}}{\partial A_{k,uv}} = \sum_{j,j'}[\Sigma_{yy|x}]_{jj'}\left(\Psi_{iu}\Psi_{vj}\bar{\Psi}_{i'j'}e^{-i\omega k} + \bar{\Psi}_{i'u}\bar{\Psi}_{vj'}\Psi_{ij}e^{i\omega k}\right), \tag{2.61}$$

so that in particular $\left.\frac{\partial T}{\partial \theta_\alpha}\right|_{\theta\in\Theta_0} = 0$ for a null parameter θ_α, and we find [*cf.* (2.59)]

$$\left.\frac{\partial^2 f_{Y\to X}}{\partial\theta_\alpha\partial\theta_\beta}\right|_{\theta\in\Theta_0} = \operatorname{tr}\left[[S_{xx}]^{-1}\left.\frac{\partial^2 T}{\partial\theta_\alpha\partial\theta_\beta}\right|_{\theta\in\Theta_0}\right] \tag{2.62}$$

for null parameters $\theta_\alpha, \theta_\beta$. From (2.61) and using (2.60), we may calculate

$$\left.\frac{\partial^2 T_{ii'}}{\partial A_{k,uv}\partial A_{k',u'v'}}\right|_{\theta\in\Theta_0} = \Psi_{iu}\Psi^*_{u'i'}\left[S_{yy|x}\right]_{vv'}e^{-i\omega(k-k')} + \Psi_{iu'}\Psi^*_{ui'}\left[S_{yy|x}\right]_{v'v}e^{i\omega(k-k')}, \tag{2.63}$$

where

$$S_{yy|x} = \Psi_{yy}\Sigma_{yy|x}\Psi^*_{yy} \tag{2.64}$$

is the CPSD for a VAR(p) model with parameters $(A_{yy}, \Sigma_{yy|x})$. From (2.62) and (2.63) we find

$$\begin{aligned}\left.\frac{\partial^2 f_{Y\to X}}{\partial A_{k,uv}\partial A_{k',u'v'}}\right|_{\theta\in\Theta_0} &= \left[[\Sigma_{xx}]^{-1}\right]_{uu'}\left\{\left[S_{yy|x}\right]_{vv'}e^{-i\omega(k-k')} + \left[S_{yy|x}\right]_{v'v}e^{i\omega(k-k')}\right\}\\ &= \left[[\Sigma_{xx}]^{-1}\right]_{uu'}\left[S_{yy|x}e^{-i\omega(k-k')} + \bar{S}_{yy|x}e^{i\omega(k-k')}\right]_{vv'}\\ &= 2\left[[\Sigma_{xx}]^{-1}\right]_{uu'}\mathfrak{Re}\left\{\left[\tilde{S}_{yy|x}\right]_{kk',vv'}\right\},\end{aligned}$$

where

$$\tilde{S}_{yy|x}(\omega) = Z(\omega)\otimes S_{yy|x}(\omega)\,, \qquad Z_{kk'}(\omega) = e^{-i\omega(k-k')}\,.$$

The $pn_y \times pn_y$ Hermitian matrix $\tilde{S}_{yy|x}(\omega)$ is the CPSD for the companion VAR(1) (Main Article, eq. 2.2) of a VAR(p) model with parameters $(A_{yy}, \Sigma_{yy|x})$, and as such may be thought of as the spectral counterpart of the autocovariance matrix $\tilde{\Gamma}_{yy|x}$ of Main Article, Section 2.4.2. Thus for $\omega \in [0, 2\pi]$, we have $W_0(\omega;\theta) = [\Sigma_{xx}]^{-1}\otimes\mathfrak{Re}\{\tilde{S}_{yy|x}(\omega)\}$, so that

$$W_0(\mathcal{F};\theta) = [\Sigma_{xx}]^{-1}\otimes\mathfrak{Re}\{\tilde{S}_{yy|x}(\mathcal{F})\}$$

with

$$\tilde{S}_{yy|x}(\mathcal{F}) = \frac{1}{|\mathcal{F}|}\int_{\mathcal{F}}\tilde{S}_{yy|x}(\omega)\,d\omega \tag{2.65}$$

as required. □

2.7.6 Proof of Main Article, Theorem 3

Lemma 1. *Suppose given a sequence of pairs of real-valued random variables* (X_n, Y_n) *such that* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{p} c$*, where* c *is a constant. Then*

$$\mathrm{pr}[X_n \leqslant Y_n] \longrightarrow \mathrm{pr}[X \leqslant c] \tag{2.66}$$

as $n \longrightarrow \infty$.

Proof. By Slutsky's Lemma [50], we have $X_n - Y_n \xrightarrow{d} X - c$. Thus for any $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that for all $n \geqslant n_0$ we have $\big|\,\mathrm{pr}[X_n - Y_n \leqslant 0] - \mathrm{pr}[X - c \leqslant 0]\,\big| < \varepsilon$, or, equivalently $\big|\,\mathrm{pr}[X_n \leqslant Y_n] - \mathrm{pr}[X \leqslant c]\,\big| < \varepsilon$, which establishes (2.66). □

of Main Article, Theorem 3. Main Article, eq. 2.38 is equivalent to $P_I(\theta;\alpha) = 1 - \mathrm{pr}\left[N\hat{F} \leqslant \Phi^{-1}_{\hat{\theta}}(1-\alpha)\right]$. Since $\hat{\theta}$ is a consistent estimator for θ and the projection of $\hat{\theta}$ onto Θ_0 is continuous—and maps any $\theta \in \Theta_0$ to itself—by the Continuous Mapping Theorem the projection of $\hat{\theta}$ onto Θ_0 converges in probability to θ. It is not hard to verify that the inverse CDF $\Phi^{-1}_{\theta}(\cdots)$ evaluated at $1-\alpha$ is continuous in the θ argument, so that again by the Continuous Mapping Theorem we have $\Phi^{-1}_{\hat{\theta}}(1-\alpha) \xrightarrow{p} \Phi^{-1}_{\theta}(1-\alpha)$. By Main Article, Theorem 1, $N\hat{F} \xrightarrow{d} Q_\theta$, where Q_θ is a (generalized χ^2) random variable with CDF Φ_θ. Applying Lemma 1 to the pair-sequence $\big(N\hat{F}, \Phi^{-1}_{\hat{\theta}}(1-\alpha)\big)$ we have

$$P_I(\theta;\alpha) \longrightarrow 1 - \mathrm{pr}\big[Q_\theta \leqslant \Phi^{-1}_{\theta}(1-\alpha)\big] = 1 - \Phi_\theta\big(\Phi^{-1}_{\theta}(1-\alpha)\big) = \alpha$$

as required. □

2.7.7 Worked example: the general bivariate VAR(1)

Consider the bivariate VAR(1)

$$X_t = a_{xx}X_{t-1} + a_{xy}Y_{t-1} + \varepsilon_{xt}\,, \tag{2.67a}$$
$$Y_t = a_{yx}X_{t-1} + a_{yy}Y_{t-1} + \varepsilon_{yt}\,, \tag{2.67b}$$

with parameters

$$A = \begin{bmatrix} a_{xx} & a_{xy} \\ a_{yx} & a_{yy} \end{bmatrix}, \qquad \Sigma = \mathrm{E}\big[\varepsilon_t \varepsilon_t^{\mathsf{T}}\big] = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix},$$

so that $\theta = (a_{xx}, a_{xy}, a_{yx}, a_{yy}, \sigma_{xx}, \sigma_{xy}, \sigma_{yy})$, and the null hypothesis H_0 (Main Article, eq. 2.17) is $a_{xy} = 0$. The transfer function is then $\Psi(\omega) = \Phi(\omega)^{-1}$ with $\Phi(\omega) = I - Az$, and the factorisation $S(\omega) = \Psi(\omega)\Sigma\Psi(\omega)^*$ (Main Article, eq. 2.8) of the CPSD $S(\omega)$ holds for $\omega \in [0, 2\pi]$.

Setting $\Delta(\omega) = |\Phi(\omega)|$ (determinant), we have

$$\Psi(\omega) = \begin{bmatrix} 1 - a_{xx}z & -a_{xy}z \\ -a_{yx}z & 1 - a_{yy}z \end{bmatrix}^{-1} = \Delta(\omega)^{-1} \begin{bmatrix} 1 - a_{yy}z & a_{xy}z \\ a_{yx}z & 1 - a_{xx}z \end{bmatrix}.$$

This leads to

$$S(\omega) = |\Delta(\omega)|^{-2} \begin{bmatrix} 1 - a_{yy}z & a_{xy}z \\ a_{yx}z & 1 - a_{xx}z \end{bmatrix} \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} \begin{bmatrix} 1 - a_{yy}\bar{z} & a_{yx}\bar{z} \\ a_{xy}\bar{z} & 1 - a_{xx}\bar{z} \end{bmatrix}$$

on $z = 1$, where $z = e^{-i\omega}$ and $\bar{z}$ its the complex conjugate[1].

We wish to calculate the Granger causality $F_{Y \to X}$. If v is the residuals variance for the VAR representation of the subprocess X_t, then the Granger causality is just $F_{Y \to X} = \log v - \log \sigma_{xx}$. To solve for v we could use the reduced DARE (2.55), but here we use an explicit spectral factorisation for the CPSD $S_{xx}(\omega)$ of X_t.

Let $\psi(\omega)$ be the transfer function of the process X_t, which satisfies the spectral factorisation $S_{xx}(\omega) = v|\psi(\omega)|^2$. We may now calculate (we denote terms we don't need by "$\cdots$").

$$\begin{aligned} S(\omega) &= |\Delta(\omega)|^{-2} \begin{bmatrix} 1 - a_{yy}z & a_{xy}z \\ \cdots & \cdots \end{bmatrix} \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} \begin{bmatrix} 1 - a_{yy}\bar{z} & \cdots \\ a_{xy}\bar{z} & \cdots \end{bmatrix} \\ &= |\Delta(\omega)|^{-2} \begin{bmatrix} 1 - a_{yy}z & a_{xy}z \\ \cdots & \cdots \end{bmatrix} \begin{bmatrix} \sigma_{xx}(1 - a_{yy}\bar{z}) + \sigma_{xy}a_{xy}\bar{z} & \cdots \\ \sigma_{yx}(1 - a_{yy}\bar{z}) + \sigma_{yy}a_{xy}\bar{z} & \cdots \end{bmatrix}. \end{aligned}$$

We now calculate (note that $z + \bar{z} = 2\cos\omega$):

$$\begin{aligned} S_{xx}(\omega) &= |\Delta(\omega)|^{-2} \{(1 - a_{yy}z)[\sigma_{xx}(1 - a_{yy}\bar{z}) + \sigma_{xy}a_{xy}\bar{z})] + a_{xy}z[\sigma_{yx}(1 - a_{yy}\bar{z}) + \sigma_{yy}a_{xy}\bar{z})]\} \\ &= |\Delta(\omega)|^{-2} \left\{\sigma_{xx}|1 - a_{yy}z|^2 + \sigma_{xy}a_{xy}[(1 - a_{yy}z)\bar{z} + (1 - a_{yy}\bar{z})z] + \sigma_{yy}a_{xy}^2 z\bar{z}\right\} \\ &= |\Delta(\omega)|^{-2} \left\{\sigma_{xx}[1 - a_{yy}(z + \bar{z}) + a_{yy}^2] + \sigma_{xy}a_{xy}(z + \bar{z} - 2a_{yy}) + \sigma_{yy}a_{xy}^2\right\} \\ &= |\Delta(\omega)|^{-2} \left\{\sigma_{xx}[1 - 2a_{yy}\cos\omega + a_{yy}^2] + 2\sigma_{xy}a_{xy}(\cos\omega - a_{yy}) + \sigma_{yy}a_{xy}^2\right\}, \end{aligned}$$

and finally

$$S_{xx}(\omega) = |\Delta(\omega)|^{-2}(P - Q\cos\omega),$$

[1] For a complex variable w, $|w|$ denotes the norm $(w\bar{w})^{\frac{1}{2}}$.

where we have set

$$P = \sigma_{xx}(1 + a_{yy}^2) - 2\sigma_{xy}a_{xy}a_{yy} + \sigma_{yy}a_{xy}^2\,, \qquad Q = 2(\sigma_{xx}a_{yy} - \sigma_{xy}a_{xy})\,.$$

The form of this expression suggests that the transfer function $\psi(\omega)$ should take the form

$$\psi(\omega) = \Delta(\omega)^{-1}(1 - bz)$$

for some constant b; this implies that X_t is VARMA(2,1). Then $|\psi(\omega)|^2 = |\Delta(\omega)|^{-2}(1 + b^2 - 2b\cos\omega)$ and the spectral factorisation $S_{xx}(\omega) = v|\psi(\omega)|^2$ now reads

$$v(1 + b^2 - 2b\cos\omega) = P - Q\cos\omega\,.$$

This must hold for all ω, so we have

$$v(1 + b^2) = P\,, \qquad vb = \tfrac{1}{2}Q\,.$$

We may now solve for v. The second equation gives $v^2b^2 = \frac{1}{4}Q^2$, so multiplying the first equation through by v we obtain the quadratic equation for v

$$v^2 - Pv + \tfrac{1}{4}Q^2 = 0\,,$$

with solutions

$$v = \tfrac{1}{2}\Big[P \pm (P^2 - Q^2)^{\frac{1}{2}}\Big].$$

We need to take the "+" solution, as this yields the correct (zero) result for the null case $a_{xy} = 0$, so that

$$F_{Y\to X} = \log\tfrac{1}{2}\Big[P + (P^2 - Q^2)^{\frac{1}{2}}\Big] - \log\sigma_{xx}\,.$$

Besides the residuals covariances Σ, only the $Y \to X$ "causal" autoregression coefficient a_{xy} and the Y autoregressive coefficient a_{yy} appear in the expression for $F_{Y\to X}$. We note that, given any $F > 0$ and a set of model parameters excluding a_{xy}, there are in general *two* possible values of a_{xy} which yield $F_{Y\to X} = F$, except in cases where no a_{xy} exists due to the stability constraint on the spectral radius, which requires that $|a_{xx}a_{yy} - a_{xy}a_{yx}| < 1$.

From Main Article, eq. 2.12, the spectral Granger causality from $Y \to X$ is

$$f_{Y\to X}(\omega) = \log(P - Q\cos\omega) - \log\left(P - Q\cos\omega - a_{xy}^2\sigma_{yy|x}\right),$$

where $\sigma_{yy|x} = \sigma_{yy} - \sigma_{xy}^2\sigma_{xx}^{-1} = \sigma_{yy}\left(1 - \kappa^2\right)$, with $\kappa = \sigma_{xy}(\sigma_{xx}\sigma_{yy})^{-\frac{1}{2}}$ the residuals correlation.

For the sampling distributions, we shall also need the (inverse of) the covariance matrix Γ_0 of the process $[X_t^\mathsf{T}\ Y_t^\mathsf{T}]^\mathsf{T}$ on the null space $a_{xy} = 0$. Solving the DLYAP equation $\Gamma_0 - A\Gamma_0 A^\mathsf{T} = \Sigma$ for

$$\Gamma_0 = \begin{bmatrix} p & r \\ r & q \end{bmatrix}$$

yields

$$\begin{aligned} p &= \left(1 - a_{xx}^2\right)^{-1} \sigma_{xx}\,, \\ r &= (1 - a_{xx}a_{yy})^{-1}\left[\sigma_{xy} + a_{xx}a_{yx}\left(1 - a_{xx}^2\right)^{-1}\sigma_{xx}\right], \\ q &= \left(1 - a_{yy}^2\right)^{-1}\left[\sigma_{yy} + 2a_{yy}a_{yx}(1 - a_{xx}a_{yy})^{-1}\sigma_{xy}\right. \\ &\qquad \left. + a_{yx}^2(1 + a_{xx}a_{yy})\left(1 - a_{xx}^2\right)^{-1}(1 - a_{xx}a_{yy})^{-1}\sigma_{xx}\right], \end{aligned}$$

and we have in particular

$$\omega_{yy} = [\Gamma_0^{-1}]_{yy} = \frac{p}{pq - r^2}\,. \tag{2.69}$$

Note also that in the null case $a_{xy} = 0$, the spectral radius is $\rho = \max\left(|a_{xx}|, |a_{yy}|\right)$.

We apply Main Article, Theorem 1 to calculate the asymptotic distribution of the single-regression estimator $\hat{F}^{\mathsf{SR}}_{Y\to X}$ on the null space. Noting that for model order $p = 1$ we have $\tilde{\Gamma} = \Gamma_0$, and setting $\Gamma_0^{-1} = [\omega_{ij}]$, we have $[\tilde{\Gamma}^{-1}]_{yy} = [\Gamma_0^{-1}]_{yy} = \omega_{yy}$, where ω_{yy} is given by (2.69). Solving the DLYAP equation (Main Article, eq. 2.32) for $\Gamma_{yy|x}$, we find that $\Gamma_{yy|x} = \left(1 - \kappa^2\right)\sigma_{yy}/(1 - a_{yy}^2)$, and the single eigenvalue of $[\tilde{\Gamma}^{-1}]_{yy}\tilde{\Gamma}_{yy|x}$ is $\lambda = \left(1 - \kappa^2\right)\sigma_{yy}\omega_{yy}/\left(1 - a_{yy}^2\right)$. By Main Article, Theorem 1 the asymptotic distribution of the single-regression estimator is thus a scaled $\chi^2(1)$:

$$N\hat{F}^{\mathsf{SR}}_{Y\to X} \xrightarrow{d} \lambda \cdot \chi^2(1) = \Gamma\left(\tfrac{1}{2}, 2\lambda\right),$$

and the Γ-approximation in this case is exact.

We also calculate the spectral Granger causality from $Y \to X$ at $\omega \in [0, 2\pi]$ as

$$f_{Y\to X}(\omega;\theta) = \log(P - Q\cos\omega) - \log\left[P - Q\cos\omega - \left(1 - \kappa^2\right)\sigma_{xx}a_{xy}^2\right].$$

We find then that

$$S_{yy|x}(\omega) = \left(1 - \kappa^2\right)\sigma_{yy}\left(1 - 2a_{yy}\cos\omega + a_{yy}^2\right)^{-1}. \tag{2.70}$$

In this case, since the model order is $p = 1$, the point-frequency null hypothesis (Main Article, eq. 2.39) coincides with the time-domain null hypothesis Main Article, eq. 2.17 (i.e., $a_{xy} = 0$), so that from Main Article, Theorem 2 we have

$$N\hat{f}_{Y\to X}(\omega) \xrightarrow{d} \lambda(\omega)\cdot\chi^2(1)\,,$$

where $\lambda(\omega) = \big(1-\kappa^2\big)\sigma_{yy}\omega_{yy}\big(1-2a_{yy}\cos\omega+a_{yy}^2\big)^{-1}$, and the asymptotic distribution for the band-limited estimator may then be calculated as per (2.65) by integrating (2.70) across the appropriate frequency range[2].

2.7.8 A random sampling scheme for VAR model parameter space

Consider, for given number of variables n and model order p, the parameter space $\Theta = \{(A,\Sigma) : A \text{ is } n\times pn \text{ with } \rho(A) < 1\,, \Sigma \text{ is } n\times n \text{ positive-definite}\}$ of VAR(p) models. Firstly, we note that the residuals covariance matrix Σ can be taken to be a *correlation* matrix; this can always be achieved by a rescaling of variables leaving Granger causalities invariant. Further Granger causality invariances under linear transformation of variables [74] allow further effective dimensional reduction of Θ; however, even under these general transformations, and under the constraint $\rho(A) < 1$, the quotient space of Θ has infinite Lebesgue measure[3]; thus we cannot generate uniform variates (it is questionable whether this would in any case be appropriate to a given empirical scenario). Here we utilize a practical and flexible scheme for generation of variates on Θ, parametrized by spectral radius ρ, log-generalized correlation[4] $\gamma = -\log|\Sigma| + \sum_i \log\Sigma_{ii}$, and population Granger causality $F = F_{Y\to X}(\theta)$, all of which have a critical impact on Granger causality sampling distributions.

To generate a random correlation matrix Σ of dimension n with given generalized correlation γ, we use the following algorithm:

1. Starting with an $n\times n$ matrix with components iid $\sim\mathcal{N}(0,1)$, compute its QR-decomposition $[Q,R]$. The matrix $M_{ij} = Q_{ij}\cdot\mathrm{sign}(R_{jj})$ is then a random orthogonal matrix.

2. Create a random n-dimensional variance vector v with components v_i iid $\sim\chi^2(1)$. The matrix $V = M\cdot\mathrm{diag}(v)\cdot M^{\mathsf{T}}$ is then positive-definite, and

[2] We may use $\int\big(1-2a\cos\omega+a^2\big)^{-1}d\omega = 2\big(1-a^2\big)^{-1}\tan^{-1}\Big(\frac{1+a}{1-a}\tan\frac{\omega}{2}\Big)$.

[3] Although the space of $n\times n$ correlation matrices has finite measure.

[4] For Gaussian covariance matrices, log-generalized correlation coincides with *multi-information* [75]. If $R = (\rho_{ij})$ is a correlation matrix with all $\rho_{ij} \ll 1$ for $i\neq j$, then $-\log|R| \approx \sum_{i<j}\rho_{ij}^2$.

for the corresponding correlation matrix $\Sigma_{ij} = V_{ij}(V_{ii}V_{jj})^{-\frac{1}{2}}$ we have $\gamma^* = -\sum_i \log v_i + \sum_i \log V_{ii}$.

If necessary, repeat steps 1,2 until $\gamma^* \geqslant \gamma$ (this may fail if γ is too large).

3. Using a binary chop, find a constant c such that, iteratively replacing $v \leftarrow v + c$, γ^* falls within an acceptable tolerance of γ (this generally converges). The correlation matrix Σ is then returned,

For a VAR coefficients matrix sequence $A = [A_1\ A_2\ \ldots\ A_p]$, the spectral radius $\rho(A)$ is given by Main Article, eq. 2.3. If λ is a constant, it is easy to show that if A' is the sequence $[\lambda A_1\ \lambda^2 A_2\ \ldots\ \lambda^p A_p]$, then $\rho(A') = \lambda\rho(A)$. Thus any VAR coefficients sequence may be exponentially weighted so that its spectral radius takes a given value. Such weighting, however, has the side-effect of exponential decay of the A_k with lag k, which is, anecdotally, unrealistic[5]. We observe empirically that we can compensate for this decay reasonably consistently across number of variables and model orders by scaling all coefficients by A_k by $\exp\left[-(pw)^{\frac{1}{2}}\right]$ for some constant w; here we choose $w = 1$, which generally achieves a more realistic gradual and approximately linear decay. To generate a random VAR model with given generalized correlation γ and given spectral radius ρ, our procedure is as follows:

1. Generate a random correlation matrix Σ with generalized correlation γ as described above.

2. Generate p $n \times n$ coefficient matrices A_k with components iid $\sim \mathcal{N}(0, 1)$. The A_k are the weighted uniformly by $\exp\left[-(pw)^{\frac{1}{2}}\right]$.

To enforce the null condition $A_{k,xy} = 0$,

3. Set all $A_{k,xy}$ components to zero.

4. Scale the A_k coefficients sequence exponentially by an appropriate constant λ, so as to achieve the given spectral radius ρ.

To instead enforce a given (non-null) population Granger causality value F,

3. Scale the $A_{k,xy}$ components uniformly by a constant c.

4. Scale the A_k coefficients sequence exponentially by an appropriate constant λ, so as to achieve the given spectral radius ρ.

Under steps $3, 4$ the population Granger causality depends monotonically on c; consequently,

[5] At least, in the authors' experience, for neural or econometric data.

5. Perform a binary chop on c, iterating steps $3, 4$ until the Granger causality is within an acceptable tolerance of F (this generally converges quickly).

In all simulations except for the bivariate model (Section 2.7.7), we used $\gamma = 1$; spectral radii and population Granger causality values are as indicated in the plots. Convergence tolerances were set to $(\text{machine } \varepsilon)^{1/2} \approx 1.5 \times 10^{-8}$ under the IEEE 754-2008 binary64 floating point standard.

2.8 Acknowledgements

Authors Gutknecht and Barnett contributed equally to this work. We would like to thank the Dr. Mortimer and Theresa Sackler Foundation, which supports the Sackler Centre for Consciousness Science. We are also grateful to Anil K. Seth and Michael Wibral for useful comments.

2.9 Author contributions

AG and LB contributed equally to this work. LB initially formulated the original research question. Subsequently, AG and LB worked in close collaboration on all other aspects of the research. They conducted simulations, derived analytical solutions, drafted the manuscript, and engaged in discussions and interpretation of results. Both authors were actively involved in revising the manuscript and have read and approved the final version for publication.

3

Introducing a differentiable measure of pointwise shared information

Abdullah Makkeh [1], Aaron J. Gutknecht [1], Michael Wibral[1]

1 Campus Institute for Dynamics of Biological Networks, Georg-August University, Goettingen, Germany

Published as: ***Makkeh, A., Gutknecht, A. J., & Wibral, M. (2021). Introducing a differentiable measure of pointwise shared information. Physical Review E, 103(3), 032149.***

Abstract

Partial information decomposition (PID) of the multivariate mutual information describes the distinct ways in which a set of source variables contains information about a target variable. The groundbreaking work of Williams and Beer has shown that this decomposition cannot be determined from classic information theory without making additional assumptions, and several candidate measures have been proposed, often drawing on principles from related fields such as decision theory. None of these measures is differentiable with respect to the underlying probability mass function. We here present a novel measure that satisfies this property, emerges solely from information-theoretic principles, and has the form of a local mutual information. We show how the measure can be understood from the perspective of exclusions of probability mass, a principle that is foundational to the original definition of the mutual information by Fano. Since our measure is well-defined for individual realizations of the random variables it lends itself for example to local learning in artificial neural networks. We also show that it has a meaningful Moebius inversion on a redundancy lattice and obeys a target chain rule. We give an operational interpretation of the measure based on the decisions that an agent should take if given only the shared information.

3.1 Introduction

What are the distinct ways in which a set of source variables may contain information about a target variable? How much information do input variables provide *uniquely* about the output, such that this information about the output variable cannot be obtained by any other input variable, or collections thereof? How much information is provided in a *shared* way, i.e., redundantly, by multiple input variables, or multiple collections of these? And how much information about the output is provided *synergistically* such that it can only be obtained by considering many or all input variables together? Answering questions of this nature is the scope of partial information decomposition (PID).

A solution to this problem has been long desired in studying complex systems [76–78] but seemed out of reach until the groundbreaking study of Williams and Beer [19]. This study provided first insights by establishing that information theory is lacking axioms to uniquely solve the PID problem. Such axioms have to be chosen in a way that satisfies our intuition about shared, unique, and synergistic information (at least in simple corner cases). However, further studies in [79, 80] quickly revealed that not all intuitively desirable properties, like positivity, zero redundant information for statistically independent input, a chain rule for composite output variables, etc., were compatible, and the initial measure proposed by Williams and Beer was rejected on the grounds of not fulfilling certain desiderata favored in the community. Nevertheless, the work of Williams and Beer clarified that indeed an axiomatic approach is necessary and also highlighted the possibility that the higher order terms (or questions) that arose when considering more than two input variables could be elegantly organized into contributions on the lattice of antichains (see more below). Approaches that do not fulfill the Williams and Beer desiderata have been suggested, e.g., [81, 82]. However, these approaches fail to quantify all the desired quantities and, therefore, answer a question different from that posed by PID.

Subsequently, multiple PID frameworks have been proposed, and each of them has merits in the application case indicated by its operational interpretation (Bertschinger et al. [83], e.g., justify their measure of unique information in a decision-theoretic setting). However, all measures lacked the property of being well defined on individual realizations of inputs and outputs (localizability), as well as continuity and differentiability in the underlying joint probability distribution. These properties are key desiderata for the settings of interest to neuroscientists and physicists, e.g., for distributed computation, where locality is needed to unfold computations in space and time [84–87]; for learning in neural networks [31, 88] where differentiability is

needed for gradient descent and localizability for learning from single samples and minibatches; for neural coding [88, 89] where localizability is important to evaluate the information value of individual inputs that are encoded by a system; and for problems from the domain of complex systems in physics as discussed in [90].

While the first two properties have very recently been provided by the pointwise partial information decomposition (PPID) of Finn and Lizier [91], differentiability is still missing, as is the extension of most measures to continuous variables. Differentiability, however, seems pivotal to exploit PID measures for learning in neural networks – as suggested for example in [31], and also in physics problems.

Therefore, we here rework the definition of Finn and Lizier [91] in order to define a novel PID measure of shared mutual information that is localizable and also differentiable. We aim for a measure that adheres as closely as possible to the original definition of (local) mutual information – in the hope that our measure will inherit most of the operational interpretation of local mutual information. We also seek to avoid invoking assumptions or desiderata from outside the scope of information theory, e.g., we explicitly seek to avoid invoking desiderata from decision or game theory. We note that adhering as closely as possible to information-theoretic concepts should also simplify finding localizable and differentiable measures.

Our goals above suggest that we have to abandon positivity for the parts (called atoms in [19]) of the decomposition, simply because the local mutual information can be already negative [1] With respect to a negative shared information in the PID we aim to preserve the interpretation of negative terms as being misinformative, in the sense that obtaining negative information will make a rational agent more likely to make the wrong prediction about the value of a target variable. Our goals also strongly suggest to avoid computing the minimum (or maximum) of multiple information expressions anywhere in the definition of the measure. This is because taking a minimum or maximum would almost certainly collide with differentiability and also a later extension to continuous variables.

The paper proceeds as follows. First, Section 3.2, introduces our measure of shared information $i_\cap^{\mathrm{sx}}$. Then, section 3.3 lays out how $i_\cap^{\mathrm{sx}}$ can be understood based on the concept of shared probability mass exclusions. Section 3.4 utilizes $i_\cap^{\mathrm{sx}}$ to obtain a full PID and establishes its differentiability. Then, Section 3.5 discusses some implications

[1] This can be seen as follows: Assuming that the negative local MI consists only of shared information, then this local shared information must be negative, enforcing the existence of negative local shared information. Now assuming that this shared information does not differ from realization to realization – something we should consider possible at this point – while the other contributions vary, then this leads to a shared information that is also negative on average, also see [91]

of $i_\cap^{\text{sx}}$ being a local mutual information, its operational interpretation, and some key applications of $i_\cap^{\text{sx}}$. Finally, Section 3.6 concludes by several examples.

3.2 Definition of the measure $i_\cap^{sx}$ of pointwise shared information

We begin by considering discrete random variables $S_1, \ldots, S_n$ and T where the S_i are called the sources and T is the target. Suppose now that these random variables have taken on particular realizations $s_1, \ldots, s_n$ and t. Our goal is to quantify the *pointwise* shared information that the source realizations carry about the target realization. We will proceed in three steps: (1) we define the information shared by *all* source realizations about the target realization, (2) we define pointwise shared information for any *subset* of source realizations, and (3) we provide the complete definition of the information shared by *multiple subsets* of source realizations.

So how much information about the target realization t is redundantly contained in all source realizations s_i? We propose that this information can be quantified as the information about the target realization provided by the truth of the statement

$$\mathcal{W}_{s_1,\ldots,s_n} = \big((S_1 = s_1) \vee \ldots \vee (S_n = s_n)\big) \tag{3.1}$$

i.e., by the inclusive OR of the statements that each source variable has taken on its specific realization. This information in turn can be understood as a regular pointwise mutual information between the target realization t and the indicator random variable [2] of the statement $\mathcal{W}_{s_1,\ldots,s_n}$ assuming the value 1:

$$i_\cap^{\text{sx}}(t : s_1; \ldots; s_n) := \log_2 \frac{p(t \mid \mathbf{I}_{\mathcal{W}_{s_1,\ldots,s_n}} = 1)}{p(t)} \tag{3.2}$$

$$= \log_2 \frac{p(t \mid \mathcal{W}_{s_1,\ldots,s_n} = \text{true})}{p(t)}. \tag{3.3}$$

The superscript "sx" stands for "shared exclusion" and will be explained in more detail in the next section. The reason for the choice of $\mathcal{W}_{s_1,\ldots,s_n}$ is the following: the truth of this statement can be verified by knowing the realization of *any* single source variable, i.e., knowing that $S_i = s_i$ for at least one i. Thus, whatever information

[2] Note that the idea of using an auxiliary random variable ($\mathbf{I}_\mathcal{W}$ in our case) is not novel per se. Quax et al. [81] has defined synergy using auxiliary random variable. However, their auxiliary random variable is conceptually different from $\mathbf{I}_\mathcal{W}$ and their approach yielded a 'stand-alone' measure of synergistic information without providing any decomposition.

can be obtained from $\mathcal{W}_{s_1,\ldots,s_n}$ can also be obtained from any individual statement $S_i = s_i$. In other words, the statement $\mathcal{W}_{s_1,\ldots,s_n}$ *only* contains information that is redundant to all source realizations. Conversely, whatever information can be obtained from all individual statements $S_i = s_i$ can also be obtained from $\mathcal{W}_{s_1,\ldots,s_n}$ because it implies that at least one of the statements $S_i = s_i$ has to be true. In other words, *all* of the information shared by the source realizations is contained in the statement $\mathcal{W}_{s_1,\ldots,s_n}$. Accordingly, the statement $\mathcal{W}_{s_1,\ldots,s_n}$ exactly captures the information redundantly contained in the source realizations. Any logically stronger or weaker statement would either contain some nonredundant information or miss out on some redundant information respectively. For a more comprehensive and foundational version of this argument, connecting principles from mereology (the study of parthood relations) and formal logic, see [92].

Now, this definition is not entirely complete yet since it only quantifies the information shared by *all* source realizations $s_1, \ldots, s_n$. However, a full-fledged measure of shared information also has to specify the information shared by (1) any *subset* of source realizations (e.g.,the information shared by s_1 and s_3) and (2) *multiple subsets* of source realizations (e.g., the information shared by (s_1, s_2) and (s_2, s_3)) [19]. The definition for a subset $\mathbf{a} \subseteq \{1, \ldots, n\}$ is straightforward: the information shared by the corresponding realizations $(s_i \mid i \in \mathbf{a})$ is the information provided by the statement

$$\mathcal{W}_{\mathbf{a}} = \left(\bigvee_{i \in \mathbf{a}} S_i = s_i \right) \tag{3.4}$$

i.e., by the logical OR of statements $S_i = s_i$ *where i is in the subset in question*. Note that in the following we will refer to sets of source realizations *by their index sets* for brevity. So we will generally say "the set of source realizations $\mathbf{a}$" instead of "the source realizations $(s_i \mid i \in \mathbf{a})$". There are formal reasons why it is preferable to work with index sets that will become apparent in Section 3.4.

Now, how about the case of multiple subsets? Note first that the pointwise mutual information provided by a given subset $\mathbf{a}$ of source realizations about the target realization is the information provided by the logical AND of the corresponding statements $S_i = s_i$:

$$i\left(t : (s_i)_{i \in \mathbf{a}}\right) = \log_2 \frac{p(t \mid (\bigwedge_{i \in \mathbf{a}} S_i = s_i) = \text{true})}{p(t)}. \tag{3.5}$$

Accordingly, the information shared by multiple subsets of source realizations $\mathbf{a}_1, \ldots, \mathbf{a}_m$ can be quantified as the information provided by the logical OR of

the associated logical AND statements, i.e., as the information provided by the statement

$$\mathcal{W}_{\mathbf{a}_1,\ldots,\mathbf{a}_m} = \left(\bigvee_{i=1}^{m} \bigwedge_{j \in \mathbf{a}_i} S_j = s_j \right). \tag{3.6}$$

The underlying reasoning is exactly as described above: whatever information can be obtained from the $\mathcal{W}_{\mathbf{a}_1,\ldots,\mathbf{a}_m}$ can also be obtained from all of the conjunctions $\bigwedge_{j \in \mathbf{a}_i} S_j = s_j$ because as soon as the truth of one of the conjunctions is known the truth of $\mathcal{W}_{\mathbf{a}_1,\ldots,\mathbf{a}_m}$ is known as well. Conversely, whatever information can be obtained from all conjunctions can also be obtained from $\mathcal{W}_{\mathbf{a}_1,\ldots,\mathbf{a}_m}$ since this statement implies that at least one conjunction must be true. This leads us to the final definition of the information shared by arbitrary subsets of source realizations $\mathbf{a}_1, \ldots, \mathbf{a}_m$:

$$i_\cap^{\mathrm{sx}}(t : \mathbf{a}_1; \ldots; \mathbf{a}_m) := \log_2 \frac{p(t \mid \mathbf{I}_{\mathcal{W}_{\mathbf{a}_1,\ldots,\mathbf{a}_n}} = 1)}{p(t)} \tag{3.7}$$

$$= \log_2 \frac{p(t \mid \mathcal{W}_{\mathbf{a}_1,\ldots,\mathbf{a}_n} = \text{true})}{p(t)}. \tag{3.8}$$

Note that this general definition agrees with the above definition of the information shared by all source realizations or subsets thereof. We would also like to emphasize here again that $i_\cap^{\mathrm{sx}}$ has the form of a local mutual information. This feature is of particular importance in the following section where we aim to provide further intuition for the measure by showing that it can also be motivated from the perspective of probability mass exclusions as discussed in [93].

3.3 Shared mutual information from shared exclusions of probability mass

Shannon information can be seen as being induced by exclusion of probability mass (e.g, [89, Sec. 2.1.3]), and the same perspective can actually be applied to the mutual information as well – as explicitly derived by Finn and Lizier [93]. In our approach to shared information, we suggest to keep intact this central information-theoretic principle that binds the exclusion of probability mass to information and mutual information. We now first review the probability exclusion perspective on local mutual information. Subsequently, we show how the measure $i_\cap^{\mathrm{sx}}$ of shared information, itself being a local mutual information, can be motivated from the same perspective as well.

3.3.1 Mutual information from exclusions of probability mass

The local mutual information [35] obtained from a realization (t, s) of two random variables T and S is

$$i(t : s) = \log_2 \frac{p(t \mid s)}{p(t)}. \tag{3.9}$$

This means that $i(t : s)$ compares the probability of observing t after observing s to the prior $p(t)$. Thus, s is said to be *informative* (resp. *misinformative*) about t if the chance of t occurring increases (resp. decreases) after observing s compared to the prior probability $p(t)$, i.e., if $i(t : s) > 0$ (resp. $i(t : s) < 0$).

The definition of $i(t : s)$ can be understood in terms of excluding certain probability mass [93] by rewriting it as

$$i(t, s) = \log_2 \frac{\mathbb{P}(\mathfrak{t}) - \mathbb{P}(\mathfrak{t} \cap \bar{\mathfrak{s}})}{1 - \mathbb{P}(\bar{\mathfrak{s}})} - \log_2 \mathbb{P}(\mathfrak{t}) \ , \tag{3.10}$$

where $\bar{\mathfrak{s}}$ is the set complement of the event $\mathfrak{s} = \{S = s\}$ and $\mathfrak{t} = \{T = t\}$. Looking at it in this way, pointwise mutual information can be conceptualized as follows (illustrated in FIG. 3.1): (i) "removing" all points from the initial sample space Ω that are incompatible with the observation of a specific s by giving them measure zero–for the event $\mathfrak{t}$ this has the consequence that a part of it is also removed, i.e., $\mathbb{P}(\mathfrak{t}) - \mathbb{P}(\mathfrak{t} \cap \bar{\mathfrak{s}})$; (ii) *rescaling* the probability measure to again have properly normalized probabilities, i.e., dividing by $1 - \mathbb{P}(\bar{\mathfrak{s}})$; and (iii) *comparing* the size of $\mathfrak{t}$ after observing s to the prior $\mathbb{P}(\mathfrak{t})$ on a logarithmic scale. The remove-rescale procedure is a conceptual way of thinking about the changes to Ω (after observing s) that are reflected in the conditional measure $\mathbb{P}(\cdot \mid \mathfrak{s})$.

This derivation of local mutual information can be generalized to any number of sources. For instance, the joint local mutual information of s_1, s_2 about t is

$$i(t : s_1, s_2) = \log_2 \frac{\mathbb{P}(\mathfrak{t}) - \mathbb{P}(\mathfrak{t} \cap (\bar{\mathfrak{s}}_1 \cup \bar{\mathfrak{s}}_2))}{1 - \mathbb{P}(\bar{\mathfrak{s}}_1 \cup \bar{\mathfrak{s}}_2)} - \log_2 \mathbb{P}(\mathfrak{t}). \tag{3.11}$$

The two conserved key principles here are that (i) the mutual information is always induced by exclusion of the probability mass related to events that are impossible after the observation of $s_1, \ldots, s_n$, i.e., $\bar{\mathfrak{s}}_1, \ldots, \bar{\mathfrak{s}}_n$, and (ii) the probabilities are rescaled by taking into account these very same exclusions. These core information-theoretic principles can be utilized to motivate the measure $i_\cap^{\mathrm{sx}}$ as explained in the next section.

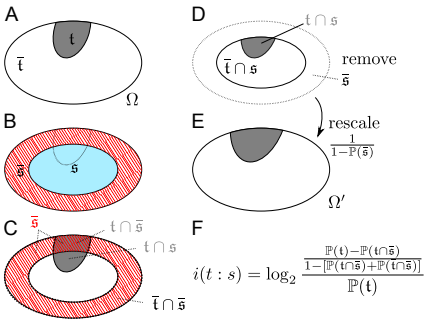

Fig. 3.1: Depiction of deriving the local mutual information $i(t:s)$ by excluding the probability mass of the impossible event $\bar{\mathfrak{s}}$ after observing $\mathfrak{s}$. (A) Two events $\mathfrak{t}, \bar{\mathfrak{t}}$ partition the sample space Ω. (B) Two event partition $\mathfrak{s}, \bar{\mathfrak{s}}$ of the source variable S in the sample space Ω. The occurrence of $\mathfrak{s}$ renders $\bar{\mathfrak{s}}$ impossible (red (dark gray) stripes). (C) $\mathfrak{t}$ may intersect with $\mathfrak{s}$ (gray region) and $\bar{\mathfrak{s}}$ (red (dark gray) hashed region). The relative size of the two intersections determines whether we obtain information or misinformation, i.e. whether $\mathfrak{t}$ becomes relatively more likely after considering $\mathfrak{s}$, or not (D), considering the necessary rescaling of the probability measure (E). Note that if the gray region in (E) is larger (resp. smaller) than that in (A), then s is informative (resp. misinformative) about t since observing s hints that t is more (reps. less) likely to occur compared to an ignorant prior. (F) shows why the misinformative exclusion $\mathbb{P}(\mathfrak{t} \cap \bar{\mathfrak{s}})$ (intersection of red (dark gray) hashes with gray region) cannot be cleanly separated from the informative exclusion, $\mathbb{P}(\bar{\mathfrak{t}} \cap \bar{\mathfrak{s}})$ (dotted outline in (C)), as stated already in [93]. This is because these overlaps appear together in a sum inside the logarithm, but this logarithm in turn guarantees the additivity of information terms. Thus the additivity of (mutual) information terms is incompatible with an additive separation of informative and misinformative exclusions *inside* the logarithms of the information measures.

3.3.2 $i_\cap^{\mathrm{sx}}$ from shared exclusions of probability mass

The core idea is now that just as mutual information is connected to the exclusion of probability mass, *shared* information should be connected to *shared* exclusions of probability mass, i.e., to possibilities being excluded redundantly by all (joint) source realizations in question. Now, what is excluded by a given joint source realization $\mathbf{a}_j$ is precisely the complement of the event $\mathfrak{a}_j = \bigcap_{i \in \mathbf{a}_j} \{S_i = s_i\}$. Thus, to evaluate the information shared by the joint source realizations $\mathbf{a}_1, \ldots, \mathbf{a}_m$, we need to remove and rescale by the intersection of the complement events $\bar{\mathfrak{a}}_j$. This intersection contains points that are excluded by all joint source realizations in question. Hence, we arrive at

$$i_\cap^{\mathrm{sx}}(t : \mathbf{a}_1; \mathbf{a}_2; \ldots; \mathbf{a}_n) := \log_2 \frac{\mathbb{P}(\mathfrak{t}) - \mathbb{P}(\mathfrak{t} \cap (\bar{\mathfrak{a}}_1 \cap \bar{\mathfrak{a}}_2 \cap \ldots \cap \bar{\mathfrak{a}}_n))}{1 - \mathbb{P}(\bar{\mathfrak{a}}_1 \cap \bar{\mathfrak{a}}_2 \cap \ldots \cap \bar{\mathfrak{a}}_n)} - \log_2 \mathbb{P}(\mathfrak{t}). \tag{3.12}$$

It is straightforward to show that this definition coincides with the one given in Section 3.2. FIG 3.2 depicts all possible exclusions in the case of three sources. This concludes our exposition of the measure of shared information $i_\cap^{\mathrm{sx}}$. In the next section, we show how this measure induces a meaningful and differentiable partial information decomposition.

3.4 Lattice structure and Differentiability

We now present a lattice structure that yields a pointwise partial information decomposition (PPID) when endowed with $i_\cap^{\mathrm{sx}}$ and show that all of the resulting PPID terms are differentiable. The lattice structure was originally introduced by Williams and Beer [19] on the basis of a range of axioms they placed on the concept of redundant information (see below). As we showed in [92] it can also be derived from elementary parthood relationships between the PID terms (also called PID atoms) and mutual information terms.

3.4.1 Lattice structure

Williams and Beer in their seminal work [19] showed that in order to capture all the information contributions that a set of sources has about a target, we need to

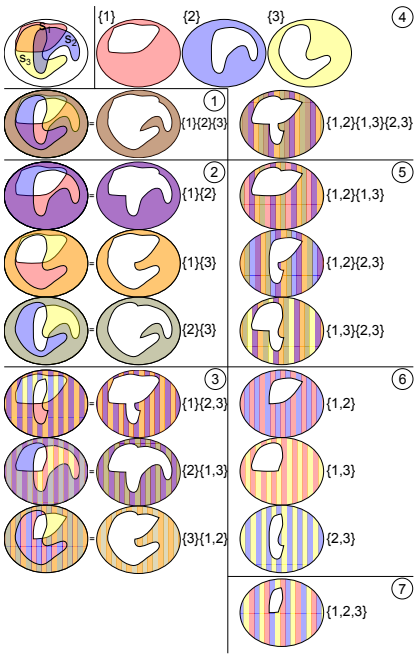

Fig. 3.2: Shared exclusions in the three-source variable case. *Upper left:* A sample space with three events $\mathfrak{s}_1$, $\mathfrak{s}_2$, $\mathfrak{s}_3$ from three source variables (their complements events are depicted in (4)). For clarity, $\mathfrak{t}$ is not shown, but may arbitrarily intersect with any intersections/unions of $\mathfrak{s}_i$. The remaining panels show the induced exclusions by different combinations of $\mathbf{a}_i$. These exclusions arise by taking the corresponding unions and intersections of sets. Which unions and intersections were taken can be deduced by the shapes of the remaining, nonexcluded regions. For (1)-(3) we show the shared exclusions for combination of singletons ((1) and (2)) and those of singletons and coalitions, such as the events of the collections (*left*) and the shared exclusions (*right*). For (4)-(7) we only show shared exclusions. The online version uses the additional, nonessential color-based mark-up of unions and intersections: An *intersection exclusion* is indicated by the *mix* of the individual colors, e.g., the $\{1\}\{2\}$ exclusion is $\bar{\mathfrak{s}}_1 \cap \bar{\mathfrak{s}}_2$ and mixes red and blue to purple, and a *union exclusion* is indicated by a *pattern* of the individual colors, e.g., the $\{1, 2\}$ exclusion is $\bar{\mathfrak{s}}_1 \cup \bar{\mathfrak{s}}_2$ and takes a red-blue pattern.

look at the level of *collections of sources*. That is, each combination of collections of sources captures a PPID term (an information contribution / information atom). Their argument was based on an analysis of the concept of redundant information, i.e., the information shared by multiple collections of sources. In particular, they argued that any measure of shared information should satisfy certain desiderata, referred to as W&B axioms (see Axioms 1, 2, and 3). These axioms imply that the domain of the shared information function can be restricted to the *antichain combinations*, i.e., any combination of collections of sources such that none of the collections is a subset of another. The reason is the following: consider collections $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, and suppose that $\mathbf{a} \subset \mathbf{b}$ (while $\mathbf{a} \not\subset \mathbf{c}$ and $\mathbf{c} \not\subset \mathbf{a}$). Then the information shared by all three collections is simply that shared by $\mathbf{a}$ and $\mathbf{c}$ since any information in $\mathbf{a}$ is automatically also contained in $\mathbf{b}$. In this way the information shared by multiple collections always reduces to the information associated with an antichain combination by removing all supersets. The measure $i_\cap^{\text{sx}}$ agrees with this result because the truth conditions of the statement $\mathcal{W}_{\mathbf{a}_1,\ldots,\mathbf{a}_m}$ are unaffected by superset removal.

Mathematically, the antichain-combinations form a lattice structure, i.e., there exists an ordering $\preceq$ of these antichain combinations such that for any pair of antichain combinations there is a unique infimum and supremum. In [19], this lattice of antichain combinations is called the *redundancy lattice* since it models inclusion of redundancies: redundant information terms associated with lower level antichains are included in redundancies associated with higher level antichains. Williams and Beer then introduced the PID terms implicitly via a Möbius Inversion over the lattice (more details in Appendix 3.7.1). We can proceed in just the same way on a pointwise level and introduce the PPID terms via a Möbius-Inversion of $i_\cap^{\text{sx}}$, i.e., via inverting the relationship

$$i_\cap^{\text{sx}}(t : \alpha) = \sum_{\beta \preceq \alpha} \pi^{\text{sx}}(t : \beta) \tag{3.13}$$

where α and β are antichain combinations. In this way each PPID term π^{sx} measures the information "increment" as we move up the lattice, i.e., the PPID term of a given node is that part of the corresponding shared information that is not already contained in any lower level shared information.

It should be mentioned at this point that the measure $i_\cap^{\text{sx}}$ actually violates one of the W&B axioms for shared information: it is not monotonically decreasing as more collections of source realizations are included. On first sight this appears to be a problem because one would expect, for instance, that the information shared by source realizations s_1, s_2 and s_3 should be *smaller than or equal to* the information

shared by s_1 and s_2. After all, the information shared by all three source realizations should be contained in the information shared by the first two. However, the violation of the monotonicity property has a natural interpretation in terms of informative and misinformative contributions to redundant information [91]: whereas each of these components individually *should* indeed satisfy the monotonicity axiom, this is not true of the total redundant information. Using the above example, the information shared by s_1, s_2, and s_3 can actually be larger than the information shared by s_1 and s_2 if the extra information in the latter shared information term (i.e., the information shared by s_1 and s_2 but *not* by s_3) is misinformative.

As shown in [93] it is possible to uniquely decompose the pointwise mutual information into an informative and a misinformative component. Since $i_\cap^{\mathrm{sx}}$ is itself a pointwise mutual information the same decomposition can be applied in order to obtain an informative pointwise shared information $i_\cap^{\mathrm{sx}+}$ (3.15a) and a misinformative pointwise shared information $i_\cap^{\mathrm{sx}-}$ (3.15b). We may then show that each of these components individually satisfies the W&B axioms. The decomposition reads

$$\begin{aligned} i_\cap^{\mathrm{sx}}(t : \mathbf{a}_1; \mathbf{a}_2; \ldots; \mathbf{a}_m) = {} & i_\cap^{\mathrm{sx}+}(t : \mathbf{a}_1; \mathbf{a}_2; \ldots; \mathbf{a}_m) \\ & - i_\cap^{\mathrm{sx}-}(t : \mathbf{a}_1; \mathbf{a}_2; \ldots; \mathbf{a}_m), \end{aligned} \tag{3.14a}$$

$$i_\cap^{\mathrm{sx}+}(t : \mathbf{a}_1; \mathbf{a}_2; \ldots; \mathbf{a}_m) := \log_2 \frac{1}{\mathbb{P}(\mathfrak{a}_1 \cup \mathfrak{a}_2 \cup \ldots \cup \mathfrak{a}_m)}, \tag{3.15a}$$

$$i_\cap^{\mathrm{sx}-}(t : \mathbf{a}_1; \mathbf{a}_2; \ldots; \mathbf{a}_m) := \log_2 \frac{\mathbb{P}(\mathfrak{t})}{\mathbb{P}(\mathfrak{t} \cap (\mathfrak{a}_1 \cup \mathfrak{a}_2 \cup \ldots \cup \mathfrak{a}_m))}. \tag{3.15b}$$

Here, the first term of (3.14a) is considered to be the informative part as it is what can be inferred from the sources (recall that $\mathbf{a}_i$ are indices of collections of sources) and we refer to it by $i_\cap^{\mathrm{sx}+}$ (3.15a). The second term of (3.14a) quantifies the (misinformative) relative loss of $p(t)$, the probability mass of the event $\mathfrak{t}$ (which actually happened) when excluding the mass of $\bar{\mathfrak{a}}_1 \cap \bar{\mathfrak{a}}_2 \cap \ldots \cap \bar{\mathfrak{a}}_n$ and we refer to it by $i_\cap^{\mathrm{sx}-}$ (3.15b).

Now, $i_\cap^{\mathrm{sx}\pm}$ should individually fulfill a pointwise version of the Williams and Beer axioms. These PPID axioms were described by Finn and Lizier [91].

Axiom 1 (Symmetry). *$i_\cap^+$ and $i_\cap^-$ are invariant under any permutation σ of collections of source events:*

$$\begin{aligned} i_\cap^+(t : \mathbf{a}_1; \mathbf{a}_2; \ldots; \mathbf{a}_m) &= i_\cap^+(t : \sigma(\mathbf{a}_1); \sigma(\mathbf{a}_2); \ldots; \sigma(\mathbf{a}_m)), \\ i_\cap^-(t : \mathbf{a}_1; \mathbf{a}_2; \ldots; \mathbf{a}_m) &= i_\cap^-(t : \sigma(\mathbf{a}_1); \sigma(\mathbf{a}_2); \ldots; \sigma(\mathbf{a}_m)). \end{aligned}$$

Axiom 2 (Monotonicity)**.** *$i_\cap^+$ and $i_\cap^-$ decreases monotonically as more source events are included,*

$$i_\cap^+(t : \mathbf{a}_1; \ldots; \mathbf{a}_m; \mathbf{a}_{m+1}) \leqslant i_\cap^+(t : \mathbf{a}_1; \ldots; \mathbf{a}_m),$$
$$i_\cap^-(t : \mathbf{a}_1; \ldots; \mathbf{a}_m; \mathbf{a}_{m+1}) \leqslant i_\cap^-(t : \mathbf{a}_1; \ldots; \mathbf{a}_m),$$

with equality if there exists $i \in [m]$ such that $\mathbf{a}_i \subseteq \mathbf{a}_{m+1}$.

Axiom 3 (Self-redundancy)**.** *$i_\cap^+$ and $i_\cap^-$ for a single source event $\mathbf{a}$ equal i^+ and i^-, respectively:*

$$i_\cap^+(t : \mathbf{a}) = h(\mathbf{a}) = i^+(t : \mathbf{a}),$$
$$i_\cap^-(t : \mathbf{a}) = h(\mathbf{a} \mid t) = i^-(t : \mathbf{a}).$$

Therefore, $i_\cap(t : \mathbf{a}) = i(t : \mathbf{a})$.

Note that $i(t : \mathbf{a}) = i^+(t; \mathbf{a}) - i^-(t; \mathbf{a})$, which is the informative–misinformative decomposition of the pointwise mutual information derived by Finn and Lizier [93]. The following theorem states that $i_\cap^{\mathrm{sx}\,\pm}$ result in a consistent PPID by showing that $i_\cap^{\mathrm{sx}\,+}$ and $i_\cap^{\mathrm{sx}\,-}$ individually fulfill the PPID axioms [91] (the proof is deferred to appendix 3.7.1.

Theorem 4. *$i_\cap^{\mathrm{sx}\,+}$ and $i_\cap^{\mathrm{sx}\,-}$ satisfy Axioms 1, 2, and 3.*

In this way the violation of monotonicity of the total shared information $i_\cap^{\mathrm{sx}}$ can be completely explained in terms of misinformative contributions. In fact, there is a another form of monotonicity that should hold as well: monotonicity over the redundancy lattice. As noted above the redundancy lattice models inclusion of redundancies. So we would expect lower level redundancies to be smaller than higher level redundancies. Again this form of monotonicity does not hold for $i_\cap^{\mathrm{sx}}$ itself but for its informative and misinformative components as expressed in the following theorem:

Theorem 5. *$i_\cap^{\mathrm{sx}\,\pm}$ increase monotonically on the redundancy lattice.*

There is another apparent problem that can be addressed using the separation into informative and misinformative components, namely, the fact that both $i_\cap^{\mathrm{sx}}$ as well as π^{sx} can be negative. This can be interpreted in terms of misinformation as well. To this end we define misinformative and informative PPID terms $\pi_\pm^{\mathrm{sx}}$ via Möbius

Inversions of $i_\cap^{\mathrm{sx}\pm}$. These informative and misinformative components of the PPID terms can be obtained recursively from $i_\cap^{\mathrm{sx}\pm}$ (see appendix 3.7.1). They stand in the relation $\pi^{\mathrm{sx}} = \pi_+^{\mathrm{sx}} - \pi_-^{\mathrm{sx}}$ to the PPID terms. Now, even though π^{sx} may be negative, its components π_+^{sx} and π_-^{sx} are non-negative.

Theorem 6. *The atoms $\pi_\pm^{\mathrm{sx}}$ are non-negative.*

In appendix 3.7.1, we will provide the necessary tools to prove the above theorems, in particular, theorem 6. To sum up, this section shows that $i_\cap^{\mathrm{sx}}$ results in a consistent and meaningful PPID. The apparent problems of violating monotonicity and non-negativity can be resolved by separating misinformative and informative components and showing that these components do satisfy the desired properties (for more discussion on the idea of misinformation within local Shannon information theory see Discussion).

This concludes our discussion of the PPID induced by the $i_\cap^{\mathrm{sx}}$. The global, variable-level PID can be obtained by simply averaging the local quantities over all possible realizations of the source and target random variables. For a complete worked example of the XOR probability distribution see Figure 3.3, subfigure H in particular. In the next section we establish the differentiability of $i_\cap^{\mathrm{sx}}$ and π^{sx}, an important advantage of these measures compared to other approaches.

3.4.2 Differentiability of $i_\cap^{\mathrm{sx}}$ and $\pi_\pm^{\mathrm{sx}}$

We will discuss the differentiability of the PPID obtained by $i_\cap^{\mathrm{sx}}$. This is a desirable property [31] that is proven to be lacking in some measures [94–96] or evidently lacking for other measures since their definitions are based on the maximum or (minimum) of multiple information quantities.

Let $\mathscr{A}([n])$ be the redundancy lattice (see section 3.7.1), $(T, S_1, \ldots, S_n)$ be discrete and finite random variables, and let us represent their joint probability distribution as a vector in $[0,1]^{|\mathcal{A}_T|\times|\mathcal{A}_{S_1}|\times\cdots\times|\mathcal{A}_{S_n}|}$. Thus, the set of all joint probability distributions of $(T, S_1, \ldots, S_n)$ forms a simplex that we denote by Δ_P. Note that $i_\cap^{\mathrm{sx}}$ and $\pi_\pm^{\mathrm{sx}}$ are functions of the probability distributions of $(T, S_1, \ldots, S_n)$ and so they can be differentiable w.r.t. the probability distributions. Formally, for a given $(T, S_1, \ldots, S_n)$, we show that $i_\cap^{\mathrm{sx}}$ and $\pi_\pm^{\mathrm{sx}}$ are differentiable over the interior of Δ_P.

Since $\log_2$ is continuously differentiable over the open domain $\mathbb{R}_+$, then using definitions (3.15a) and (3.15b), $i_\cap^{\mathrm{sx}+}$ and $i_\cap^{\mathrm{sx}-}$ are both continuously differentiable over the interior of Δ_p. Now, for $\alpha \in \mathscr{A}([n])$, using theorem 7 and proposition 7

$$\pi_+^{\mathrm{sx}}(t:\alpha) = \sum_{\gamma \in \mathscr{P}(\alpha^- \setminus \{\gamma_1\})} (-1)^{|\gamma|} \log_2 \left(\frac{p(\gamma) + d_1}{p(\gamma)} \right), \tag{3.16}$$

where $\alpha^- = \{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ are the children of α ordered increasingly w.r.t. their probability mass and $\alpha^- := \{\beta \in \mathscr{A}([n]) \mid \beta \prec \alpha, \beta \preceq \gamma \prec \alpha \Rightarrow \beta = \gamma\}$. Hence, π_+^{sx} is continuously differentiable over the interior of Δ_P since the function $x+d_1/x$ and its inverse are continuously differentiable over the open domain $\mathbb{R}_+$. Similarly, π_-^{sx} is continuously differentiable over the interior Δ_P.

3.5 Discussion

In this section, we first present further properties of $i_\cap^{\mathrm{sx}}$. Then, we provide an operational interpretation of $i_\cap^{\mathrm{sx}}$, and suggest an approach to compare this operational interpretation with that of other measures. Following this, we give the intuition behind the "intrinsic dependence" of PID atoms for joint source-target distributions where the number of these atoms is larger than these distributions' alphabet size. Finally, we provide two applications where $i_\cap^{\mathrm{sx}}$ is particularly well suited and discuss the computational complexity of $i_\cap^{\mathrm{sx}}$.

3.5.1 Direct consequences of $i_\cap^{\mathrm{sx}}$ being a local mutual information

The fact that $i_\cap^{\mathrm{sx}}$ has the form of a regular local mutual information has several interesting consequences.

Implied entropy decomposition Since the local entropy of a realization of a set of variables can be written as a self-mutual information our decomposition also directly implies an entropy decomposition that inherits the properties of the lattices described in section 3.4. We start by the local entropy $h(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ of a set of collections of realizations of variables $S_i = s_i$. Note that these collections have to be considered jointly, hence the comma [3]. Thus, we can equally well write the entropy that is to be decomposed as $h(\{s_i \mid i \in \bigcup \mathbf{a}_j\})$. Thus, we can consider the

[3] If the collections where considered in an OR relation, there would be no random variable on which the average entropy is defined (see discussion of the local indicator variable $w_{\mathbf{a}_1,\ldots,\mathbf{a}_m}$)

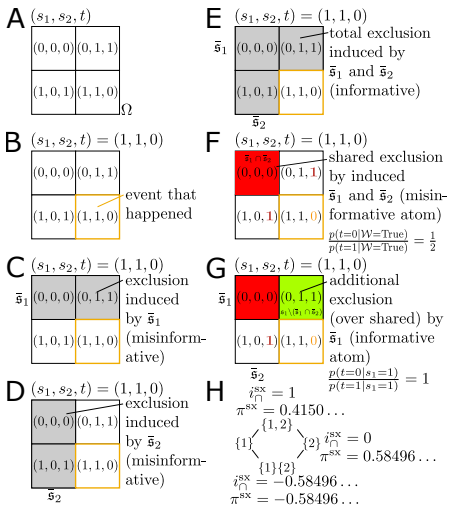

Fig. 3.3: Worked example of $i_\cap^{\text{sx}}$ for the classical XOR. Let $T = XOR(S_1, S_2)$ and $S_1, S_2 \in \{0, 1\}$ be independent uniformly distributed and consider the realization $(s_1, s_2, t) = (1, 1, 0)$. (A-B) The sample space Ω and the realized event (gold (gray) frame). (C) The exclusion of events induced by learning that $S_1 = 1$, i.e. $\bar{\mathfrak{s}}_1 = \{0\}$ (gray). (D) Same for $\bar{\mathfrak{s}}_2 = \{0\}$. (E) The union of exclusions fully determines the event $(1, 1, 0)$ and yields 1 bit of $i(t = 0 : s_1 = 1, s_2 = 1)$. (F) The shared exclusions by $\bar{\mathfrak{s}}_1 = \{0\}$ and $\bar{\mathfrak{s}}_2 = \{0\}$, i.e., $\bar{\mathfrak{s}}_1 \cap \bar{\mathfrak{s}}_2$ exclude only $(0, 0, 0)$. This is a misinformative exclusion, as it raises the probability of events that did not happen $(t = 1)$ relative to those that did happen $(t = 0)$ compared to the case of complete ignorance. (G) Learning about one full variable, i.e., obtaining the statement that $\bar{\mathfrak{s}}_1 = \{0\}$ adds additional probability mass to the exclusion (green (light gray)). The shared exclusion (red (dark gray)) and the additional unique exclusion (green (light gray)) induced by s_1 create an exclusion that is uninformative, i.e., the probabilities for $t = 0$ and $t = 1$ remain unchanged by learning $s_1 = 1$. At the level of the π^{sx} atoms, the shared and the unique information atom cancel each other. (H) Lattice with $i_\cap^{\text{sx}}$ and π^{sx} terms for this realization. Other realizations are equivalent by the symmetry of XOR, thus, the averages yield the same numbers. Note that the necessity to cancel the negative shared information twice to obtain both $i(t = 0 : s_1 = 1) = 0$ and $i(t = 0 : s_2 = 1) = 0$, results in a synergy < 1 bit. Also note that while adding the shared exclusion from (F) and the unique exclusions for s_1 and s_2 results in the full exclusion from (E), information atoms add differently due to the nonlinear transformation of excluded probability mass into information via $-\log_2 p(\cdot)$ – compare (H).

s_i together as a joint random variable whose entropy is to be decomposed. This can be done by realizing first $h(\{s_i \mid i \in \bigcup \mathbf{a}_j\}) = i(\{s_i \mid i \in \bigcup \mathbf{a}_j\} : \{s_i \mid i \in \bigcup \mathbf{a}_j\})$, and then applying our PID formalism. In this decomposition then terms of the form $i_\cap^{\mathrm{sx}}(\{s_i \mid i \in \bigcup \mathbf{a}_j\} : \mathbf{a}_1; \ldots; \mathbf{a}_m) =: h_\cap^{\mathrm{sx}}(\mathbf{a}_1; \ldots; \mathbf{a}_m)$ appear. In other words, on the target side of the arguments of $i_\cap^{\mathrm{sx}}$ we will always find the joint random variable, whereas the collections appear as usual on the source side.

Target chain rule and average measures Another consequence is that $i_\cap^{\mathrm{sx}}$ satisfies a target chain rule for a composite target variable $T = \{t_1, t_2\}$:

$$\begin{aligned} i_\cap^{\mathrm{sx}}(t_1, t_2 : \mathbf{a}_1; \mathbf{a}_2; \ldots; \mathbf{a}_m) = i_\cap^{\mathrm{sx}}(t_1 : \mathbf{a}_1; \mathbf{a}_2; \ldots; \mathbf{a}_m) \\ + i_\cap^{\mathrm{sx}}(t_2 : \mathbf{a}_1; \mathbf{a}_2; \ldots; \mathbf{a}_m \mid t_1), \end{aligned}$$

where the second term is $\log_2 \frac{\mathbb{P}(\mathrm{t}_2|\mathrm{t}_1) - \mathbb{P}(\mathrm{t}_2, \bar{\mathrm{a}}_1, \bar{\mathrm{a}}_2, \ldots, \bar{\mathrm{a}}_m|\mathrm{t}_1)}{1 - \mathbb{P}(\bar{\mathrm{a}}_1, \bar{\mathrm{a}}_2, \ldots, \bar{\mathrm{a}}_m|\mathrm{t}_1)} - \log_2 \mathbb{P}(\mathrm{t}_2 \mid \mathrm{t}_1)$. Moreover, by linearity of the averaging a corresponding target chain rule is satisfied for the average shared information, $I_\cap^{\mathrm{sx}}$, defined by.

$$\begin{aligned} I_\cap^{\mathrm{sx}}(T : \mathbf{A}_1; \ldots; \mathbf{A}_m) &:= \sum_{t, s_1, \ldots, s_n} p(t, s_1, \ldots, s_n) i_\cap^{\mathrm{sx}}(t : \mathbf{a}_1; \ldots; \mathbf{a}_m) \\ &= \sum_{t, s_1, \ldots, s_n} p(t, \mathbf{s}_1, \ldots, \mathbf{s}_n) i(t : W_{\mathbf{a}_1, \ldots, \mathbf{a}_m} = 1), \end{aligned} \tag{3.17}$$

where probabilities related to the indicator variable $W_{\mathbf{a}_1, \ldots, \mathbf{a}_m}$ have to be recomputed for each possible combination of source and target realizations. Note that this indicator variable simply indicates the truth of the statement $\mathcal{W}_{\mathbf{a}_1, \ldots, \mathbf{a}_m}$ from section 3.2. Also note that in Eq. (3.17) the averaging still runs over all combinations of $t, s_1, \ldots, s_n$, and the weights are still given by $p(t, s_1, \ldots, s_n)$, not $p(t, W_{\mathbf{a}_1, \ldots, \mathbf{a}_m} = 1)$. Having different variables in the averaging weights and the local mutual information terms makes the average shared information structurally different from a mutual information [4]. One consequence of this is that in principle the average $I_\cap^{\mathrm{sx}}$ can be negative. This also holds for the averages of the other information atoms on the lattice (see next section for the lattice structure). Thus, the local shared information may be expressed as a local mutual information with an auxiliary variable constructed for that purpose, and multiple such variables have to be constructed for a definition of a global shared information.

[4] As was to be expected from the difficulties encountered in the past trying to define measures of shared information.

Upper bounds. First, we can assess the self-shared information of a collection of variables:

$$\begin{aligned} i_{\cap}^{\mathrm{sx}}(\mathbf{a}_1;\ldots;\mathbf{a}_m : \mathbf{a}_1;\ldots;\mathbf{a}_m) &:= i(W_{\mathbf{a}_1,\ldots,\mathbf{a}_m} = 1 : W_{\mathbf{a}_1,\ldots,\mathbf{a}_m} = 1) \\ &= h(W_{\mathbf{a}_1,\ldots,\mathbf{a}_m} = 1)\ , \end{aligned} \tag{3.18}$$

where the notation $\mathbf{a}_1;\ldots;\mathbf{a}_m$ means the event defined by the complement of the intersection of exclusions induced by the $\mathbf{a}_i$, as before. This quantity is greater than or equal to zero and is the upper bound of shared information that the source variables can have about any realization u of any target variable U, i.e.,

$$i_{\cap}^{\mathrm{sx}}(\mathbf{a}_1;\mathbf{a}_2;\ldots;\mathbf{a}_m : \mathbf{a}_1;\mathbf{a}_2;\ldots;\mathbf{a}_m) \geqslant i_{\cap}^{\mathrm{sx}}(u : \mathbf{a}_1;\mathbf{a}_2;\ldots;\mathbf{a}_m)$$

for any $u \in \mathcal{A}_U$. This upper bound has conceptual links to maximum extractable shared information from [97]. Moreover, this upper bound may be nonzero even for independent sources, showing how the so-called mechanistic shared information arises.

3.5.2 Operational interpretation of $i_{\cap}^{\mathrm{sx}}$

Being a local mutual information, $i_{\cap}^{\mathrm{sx}}$ keeps all the operational interpretations of that measure. For example, in keeping with Woodward [98] it measures the information available in the statement $\mathcal{W}$ for inference about the value t of the target. Specifically, a negative value of the local shared information indicates that an agent who is only in possession of the shared information is more likely to mispredict the outcome of the target (e.g., FIGs 3.3, 3.4) than without the shared information; a positive value means that the shared information makes the agent more likely to choose the correct outcome. The unsigned magnitude of the shared information informs us about how relatively certain the agent should be about their prediction.

What remains to be clarified then is the meaning of the average expression $I_{\cap}^{\mathrm{sx}}$. As detailed above the average is taken with respect to the probabilities of the realizations of the source variables and the target variable, not with respect to the dummy variables encoding the truth value of the respective statements $\mathcal{W}$ — as an average mutual information would require. To understand the meaning of this particular average it is instructive to start by ruling out two false interpretations. Again, consider an agent who tries to predict the correct value of target t. In order to do so, the agent utilizes a particular information channel.

For the first false interpretation, consider a channel that takes the realizations of sources and target and produces the statements $\mathcal{W}$ carrying the shared information.

If the receiver of this channel used it multiple times in the case of a negative $I_\cap^{\mathrm{sx}}$, then this receiver would *learn* that the shared information received is negative on average and could modify their judgment. This leads us to a second false interpretation: the average could be understood as an average over an *ensemble* of agents, where each agent uses the above channel only once, thus avoiding the issue just described. Even in this scenario however there is a problem: if the agent knew that the information provided by $\mathcal{W}$ is shared by the true source realizations, then the agent could derive the truth of all sub-statements of $\mathcal{W}$. Accordingly, the agents would receive more than only the shared information.

In order to obtain the appropriate interpretation of shared information we have to consider a channel that masks the metainformation that all substatements of $\mathcal{W}$ are true, and also makes learning impossible. This is achieved by a channel that produces true statements $\mathcal{V}$ about the source variables which have the logical structure of $\mathcal{W}$, but do not always carry shared information. Consider the information shared by all sources. In this case the channel would randomly produce (true) statements of the form $\mathcal{V}_{s_1,\ldots,s_n} = \big((S_1 = s_1) \vee \ldots \vee (S_n = s_n)\big)$ but where some of the substatements might be false. Then $\mathcal{V}$ does not always carry shared information (only in case all substatements happen to be true). The receiver knows the joint distribution of sources and target and performs inference on t in a Bayes optimal way. Such a channel would provide non-negative average mutual information. However, for a channel of this kind the average taken to compute $I_\cap^{\mathrm{sx}}$, is only over those channel uses where $\mathcal{V}$ actually did encode shared information. In certain cases this average can be negative (see Table 3.1).

As already alluded to above, the setting of our operational interpretation contrasts with that of other approaches to PID that take the perspective of multiple agents having *full* access to individual source variables (or collections thereof), and that then design measures of unique and redundant information based on actions these agents can take or rewards they obtain in decision- or game-theoretic settings based on their access to full source variables (e.g., in [83, 91, 96]). While certainly useful in the scenarios invoked in [83, 91, 96], we feel that these operational interpretations may almost inevitably mix inference problems (i.e., information theory proper) with decision theory. Also, they typically bring with them the use of minimization or maximization operations to satisfy the competitive settings of decision or game theory. This, in turn, renders it difficult to obtain a differentiable measure of local shared information.

In sum, we feel that the question of how to decompose the information provided by multiple source variables about a target variable may indeed not be a single

Tab. 3.1: $\mathcal{V}$**-channel for XOR.** *Left:* probability masses for each realization. *Middle:* Equiprobable $\mathcal{V}$-statements associated with each realization such that respective statement carrying shared information is listed first (marked by $\mathcal{W}$) *Right:* predicted target inferred from $\mathcal{V}$ and where ✓refers to *correct* predictions and ✗ refers to *incorrect* ones. Using $\mathcal{V}$ a receiver obtains positive average mutual information, but the contribution of $\mathcal{W}$ statements is negative. *Bottom:* the sign of $I^{\mathcal{V}}$, the average information provided by all $\mathcal{V}$-statements, and that of $I_{\cap}^{\mathrm{sx}}$.

	Realization			Channel Output	Inference	
p	s_1	s_2	t	$\mathcal{V}$-statement	predicted t	Correct?
1/4	0	0	0	$(S_1 = 0) \vee (S_2 = 0)$ $(\mathcal{W})$	1	✗
				$(S_1 = 0) \vee (S_2 = 1)$	0	✓
				$(S_1 = 1) \vee (S_2 = 0)$	0	✓
1/4	0	1	1	$(S_1 = 0) \vee (S_2 = 1)$ $(\mathcal{W})$	0	✗
				$(S_1 = 0) \vee (S_2 = 0)$	1	✓
				$(S_1 = 1) \vee (S_2 = 1)$	1	✓
1/4	1	0	1	$(S_1 = 1) \vee (S_2 = 0)$ $(\mathcal{W})$	0	✗
				$(S_1 = 1) \vee (S_2 = 1)$	1	✓
				$(S_1 = 0) \vee (S_2 = 0)$	1	✓
1/4	1	1	0	$(S_1 = 1) \vee (S_2 = 1)$ $(\mathcal{W})$	1	✗
				$(S_1 = 1) \vee (S_2 = 0)$	0	✓
				$(S_1 = 0) \vee (S_2 = 1)$	0	✓

$I^{\mathcal{V}}(T : S_1; S_2) > 0$ (4 ✗ and 8 ✓) and $I_{\cap}^{\mathrm{sx}}(T : S_1; S_2) < 0$ (4 ✗ and 0 ✓)

question, but multiple questions in disguise. The most useful answer will therefore depend on the scenario where the question arose. Our answer seems to be useful in communication settings, and where quantitative statements about dependencies between variables are important (e.g., the field of statistical inference, where the PID enumerates all possible types of dependencies of the dependent (target) variable on the independent (source) variables).

3.5.3 Evaluation of $I_{\cap}^{\mathrm{sx}}$ on P and on optimization distributions obtained in other frameworks.

Since our approach to PID relies only on the original joint distribution P it can be applied to other PID frameworks where distributions $Q(P)$ are derived from the original P of the problem – e.g., via optimization procedures, as it is done for example in [83, 96]. This yields some additional insights into the operational interpretation of our approach compared to others, by highlighting how the optimization from P to $Q(P)$ shifts information between PID atoms in our framework.

3.5.4 Number of PID atoms vs alphabet size of the joint distribution

The number of lattice nodes rises very rapidly with increasing numbers of sources. Thus, the number of lattice nodes may outgrow the joint symbol count of the random variables, i.e., the number of entries in the joint probability distribution. One may ask, therefore, about the independence of the atoms on the lattice in those cases (remember that the atoms were introduced in order to have the "independent" information contributions of respective variable configurations at the lattice nodes). As shown in Fig. 3.5 and 3.6 our framework reveals multiple additional constraints at the level of exclusions via the family of mappings from Proposition 7. This explains mechanistically why not all atoms are independent in cases where the number of atoms is larger than the number of symbols in the joint distribution.

3.5.5 Key applications

Due to the fact that PID solves a basic information-theoretic problem, its applications seem to cover almost all fields where information theory can be applied. Here, we focus on two applications for which our measure is suited particularly well: the first application requires localizability and differentiability; the second application does not require differentiability, but requires at least continuity of the measure on the space of the underlying probability distributions.

Learning neural goal functions

In [31] we argued that information theory, and in particular the PID framework, lends itself to unify various neural goal functions, e.g., infomax and others. We also showed how to apply this to learning in neural networks via the coherent infomax framework of Kay and Phillips [88]. Yet, this framework was restricted to goal functions expressible using combinations, albeit complex ones, of terms from classic information theory, due to the lack of a differentiable PID measure. Goal functions that were only expressible using PID proper could not be learned in the Kay and Phillips framework, and in those cases PID would only serve to assess the approximation loss.

Our new measure removes this obstacle and neural networks or even individual neurons can now be devised to learn pure PID goal functions. A possible key application is in hierarchical neural networks with a hierarchy of modules, where each module contains two populations of neurons. These two populations represent supra-

and infragranular neurons and coarsely mimic their different functional roles. One population represents so-called layer 5 pyramidal cells. It serves to send the shared information between their bottom-up (e.g., sensory) inputs and their top-down (contextual) inputs downwards in the hierarchy; the other population represents layer 3 pyramidal cells and sends the synergy between the bottom-up inputs and the top-down inputs upwards in the hierarchy. For the first population the extraction of shared information between higher and lower levels in the hierarchy can be roughly equated to learning an internal model, while for the second population the extraction of synergy is akin to computing a generalized error (see [99, 100] and references therein for the neuroanatomic background of this idea). Thus, a hierarchical network of this kind can perform an elementary type of predictive coding. The full details of this application scenario are the topic of another study, however.

Information modification in distributed computation in complex systems

If one desires to frame distributed computation in complex systems in terms of the elementary operations on information performed by a Turing machine, i.e., the storage, transfer, and modification of information, information-theoretic measures for each of these component operations are required. For storage and transfer well established measures are available, i.e., the active information storage [84] and the transfer entropy [85–87]. For modification, in contrast, no established measures exist, yet an appropriate measure of synergistic mutual information from a partial information decomposition has been proposed as a candidate measure of information modification [101]. An appropriate measure in this context has to be localizable (i.e., it must be possible to evaluate the measure for a single event) in order to serve as an analysis of computation locally in space and time, and it has to be continuous in terms of the underlying probability distribution. Both of these conditions were already met for the PPID measure of Finn and Lizier [91]; our novel measure here adds the possibility to differentiate the measure on the interior of the probability simplex, which makes it even more like a classic information measure. This is important to determine the input distribution that maximizes synergy in a system, i.e., the input distribution that reveals the information modification capacity of the computational mechanism in a system as suggested in [102].

3.5.6 Computational complexity of the PID using $i_\cap^{\mathrm{sx}}$

Real-world applications of PID will not necessarily be confined to the standard two-input variable case – hence the importance of the organization scheme for higher order terms that are provided by the lattice structure. For such real-world problems the computational complexity of the computation of each atom on the lattice becomes important – not least because of the potentially large number of atoms (see below). This holds in particular when additional nonparametric statistical tests of PID measures obtained from data require many recomputations of the measures. We, therefore, discuss the computational complexity of our approach.

For each realization $s = (s_1, \ldots, s_n)$ and t, our PPID is obtained by computing the atoms $\pi_\pm^{\mathrm{sx}}(t : \alpha)$ for each $\alpha \in \mathscr{A}([n])$. In Appendix 3.7.1, we show that any $\pi_\pm^{\mathrm{sx}}(t : \alpha)$ is evaluated as follows:

$$\pi_\pm^{\mathrm{sx}}(t : \alpha) = i_\cap^{\mathrm{sx}\,\pm}(t : \alpha) - \sum_{\beta \prec \alpha} \pi_\pm^{\mathrm{sx}}(t : \beta) \quad \forall\, \alpha, \beta \in \mathscr{A}([n]),$$

where computing any $i_\cap^{\mathrm{sx}}(t : \alpha)$ is linear in the size of $\mathcal{A}_{T,S}$, the alphabet of the joint random variable $(T, S_1, \ldots, S_n)$. Moreover, using $i_\cap^{\mathrm{sx}}$ as a redundancy measure, the closed form of $\pi_\pm^{\mathrm{sx}}$ derived in (3.16) shows that the computation of our PID is trivially parallelizable over atoms and realizations, which is crucial for larger number of sources. The importance of parallelization is due to the rapid growth of PID terms M when the number of sources gets larger for any PID lattice-based measure. This M grows super exponentially as the n-th Dedekind number $d(n) - 2$. At present even enumerating M is practically intractable beyond $n > 8$.

3.6 Examples

In this section, we present the PID provided by our $i_\cap^{\mathrm{sx}}$ measure for some exemplary probability distributions. Most of the distributions are chosen from Finn and Lizier [91] and previous examples in the PID literature. The code for computing π^{sx} is available on the IDTxl toolbox `http://github.com/pwollstadt/IDTxl` [103].

3.6.1 Probability distribution PwUnq

We start by the pointwise unique distribution (PwUnq) introduced by Finn and Lizier [91]. This distribution is constructed such that for each realization, only one of the sources holds complete information about the target while the other holds no

Tab. 3.2: PWUNQ Example. *Left:* probability mass diagrams for each realization. *Right:* the pointwise partial information decomposition for the informative and misinformative. *Bottom:* the average partial information decomposition.

	Realization			π_+^{sx}				π_-^{sx}			
p	s_1	s_2	t	$\{1\}\{2\}$	$\{1\}$	$\{2\}$	$\{1,2\}$	$\{1\}\{2\}$	$\{1\}$	$\{2\}$	$\{1,2\}$
1/4	0	1	1	1	0	1	0	1	0	0	0
1/4	1	0	1	1	1	0	0	1	0	0	0
1/4	0	2	2	1	0	1	0	1	0	0	0
1/4	2	0	2	1	1	0	0	1	0	0	0
Average Values				1	1/2	1/2	0	1	0	0	0

$\Pi^{\mathrm{sx}}(T:\{1\}\{2\}) = 0$ $\Pi^{\mathrm{sx}}(T:\{1\}) = 1/2$ $\Pi^{\mathrm{sx}}(T:\{2\}) = 1/2$ $\Pi^{\mathrm{sx}}(T:\{1,2\}) = 0$

information. The aim was to structure a distribution where at no point (realization) the two sources give the same information about the target. Hence, Finn and Lizier argue that, for such distribution, there should be no shared information. Also, this distribution highlights the need for a pointwise analysis of the PID problem.

Since in all of the realizations, the shared exclusion does not alter the likelihood of any of the target events compared to the case of total ignorance, $i_\cap^{\mathrm{sx}}$ will indeed give zero redundant information. Thus, the PID terms resulting from $i_\cap^{\mathrm{sx}}$ are the same as the those resultant from $r_{\min}$ [91] and I_{ccs} [96] measures (see table 3.2).

Recall Assumption $(*)$ of Bertschinger et al. [83] which states that the unique and shared information should only depend on the marginal distributions $P(S_1, T)$ and $P(S_2, T)$. Finn and Lizier [91] showed that all measures which satisfy Assumption $(*)$ result in no unique information, i.e., nonzero redundant information whenever $P(S_1, T)$ is isomorphic to $P(S_2, T)$. The PWUNQ distribution falls into this category for which $I_{\min}$ [19], I_{red} [80], $\widetilde{\mathrm{UI}}$ [83], and $\mathcal{S}_{\mathrm{VK}}$ [104] do not register unique information of S_1 and S_2. This is due to Assumption $(*)$ not taking into consideration the pointwise nature of information. Specifically, a measure that satisfies Assumption $(*)$ is agnostic to the fact that at each realization $\{T = j\}$ is uniquely determined by S_1 or S_2 but never both. On the contrary such a measure registers this as a mixture of shared and synergistic contribution since neither S_1 nor S_2 can fully determine $\{T = j\}$ on its own but shared they partly determine $\{T = j\}$.

3.6.2 Probability distribution XOR

Using our formulation of $i_\cap^{\mathrm{sx}}$ results in negative local shared information for the classic XOR example. To see this, assume that S_1 and S_2 are independent, uni-

formly distributed random bits, and $T = \mathrm{XOR}(S_1, S_2)$, and consider the realization $(s_1, s_2, t) = (1, 1, 0)$. From Eq. (3.12) we get

$$i_\cap^{\mathrm{sx}}(t = 0 : s_1 = 1; s_2 = 1) = \log_2 \frac{1/2 - 1/4}{1 - 1/4} + \log_2 \frac{1}{1/2} < 0.$$

We argue that this result reflects that an agent receiving the shared information is misinformed (see, e.g., [93] for the concept of misinformation) about t. To understand the source of this misinformation, consider that the agent is only provided with the shared information, i.e., the agent knows only that $\mathcal{W}_{s_1,\ldots,s_n}$ is true. This means the agent is being told the following: "One of the two sources has outcome 1, and we do not know which one." This will let the agent predict that the joint realization is one out of three realizations with equal probability: $(1, 1, 0)$, $(0, 1, 1)$, or $(1, 0, 1)$ (see FIG 3.3). Of these three realizations, only one points to the correct target realization $t = 0$, while the other two point to the "wrong" $t = 1$ leading to odds of 1:2 — whereas $t = 0$ and $t = 1$ were equally probable before the agent received the shared information from the sources. As a consequence, the local shared information becomes negative [5]. Finally, the XOR gate demonstrates an example of negative shared information; we note that in general unique (e.g., table 3.3) and synergistic information can as well be negative.

3.6.3 Probability distribution RNDERR

Recall RND, the redundant probability distribution, where both sources are fully informative about the target and exhibit the same information. More precisely, the *redundant realizations*, $s_1 = s_2 = t = 0$ and $s_1 = s_2 = t = 1$, are the only two realizations that occur equally likely. Derived from RND, the RNDERR is a noisy redundant distribution of two sources where one source occasionally misinforms about the target while the other remains fully informative about the target. Moreover, if S_2 is the source that occasionally misinforms about the target, then the *faulty realizations*, namely, $s_2 \neq s_1 = t = 0$ and $s_2 \neq s_1 = t = 1$, are equally likely, but less likely than the redundant ones. We stick to the probability masses given in [91] for

[5] Due to $i(t : s_j) = 0$ for $j = 1, 2$ in the XOR example, this negative shared information is then compensated by positive unique information – however this happens twice, i.e. once for each marginal local mutual information. As a consequence, the synergy is reduced from 1 bit to 1 minus once this unique information. This may seem counter-intuitive when still thinking about the PID atoms as areas, in the sense of "How come if we subtract two mutual information of zero bit from the joint mutual information of 1 bit, that we do not get 1 bit as a result?". The key insight is that the two local mutual information terms of zero bit have a negative "overlap" with each other, making their sum positive. We simply see here again that the interpretation of PID atoms as (semi-positive) areas has to be given up in the pointwise framework, due to the fact that already the regular local mutual information can be negative.

Tab. 3.3: RndErr Example. *Left:* probability mass diagrams for each realization. *Right:* the pointwise partial information decomposition for the informative and misinformative is evaluated. *Bottom:* the average partial information decomposition. We set $a = \log_2(8/5), b = \log_2(8/7), c = \log_2(5/4), d = \log_2(7/4), e = \log_2(16/15), f = \log_2(16/17)$, and $g = \log_2(4/3)$.

	Realization			π_+^{sx}				π_-^{sx}			
p	s_1	s_2	t	$\{1\}\{2\}$	$\{1\}$	$\{2\}$	$\{1,2\}$	$\{1\}\{2\}$	$\{1\}$	$\{2\}$	$\{1,2\}$
3/8	0	0	0	a	c	c	e	0	0	g	0
3/8	1	1	1	a	c	c	e	0	0	g	0
1/8	0	1	0	b	d	d	f	0	0	2	0
1/8	1	0	1	b	d	d	f	0	0	2	0
Average Values				0.557	0.443	0.443	0.367	0	0	0.811	0

$\Pi^{\mathrm{sx}}(T:\{1\}\{2\}) = 0.557\ \Pi^{\mathrm{sx}}(T:\{1\}) = 0.443\ \Pi^{\mathrm{sx}}(T:\{2\}) = -0.367\ \Pi^{\mathrm{sx}}(T:\{1,2\}) = 0.367$

the redundant realizations 3/8 and for the faulty realizations 1/8 and speculate that S_2 will hold misinformative (negative) unique information about T.

For this distribution, our measure results in the following PID: misinformative unique information by S_2, informative unique information by S_1, informative shared information, and informative synergistic information that balances the misinformation of S_2 (see table 3.3).

3.6.4 Probability distribution XorDuplicate

In this distribution, we extend the Xor distribution by adding a third source S_3 such that (i) S_3 is a copy of any of the two original sources and (ii) S_3 does not have an additional effect on the target, e.g., if S_3 is a copy of S_1 then $T := \mathrm{Xor}(S_1, S_2) = \mathrm{Xor}(S_2, S_3)$. Let S_1 and S_2 be two independent, uniformly distributed random bits, S_3 be a copy of S_1, and $T = \mathrm{Xor}(S_1, S_2)$. This distribution (S_1, S_2, S_3, T) is called XorDuplicate where the only nonzero realizations are $(0,0,0,0), (0,1,0,1), (1,0,1,1), (1,1,1,0)$.

The key point is that the target T in the classical Xor is specified only by (S_1, S_2), whereas in XorDuplicate the target is equally specified by the coalitions (S_1, S_2) and (S_2, S_3). This means that the synergy $\Pi^{\mathrm{sx}}(T:\{1,2\})$ in Xor should be captured by the term $\Pi^{\mathrm{sx}}(T:\{1,2\}\{2,3\})$ in XorDuplicate.

The XorDuplicate distribution was suggested by Griffith et al. [104]. The authors speculated that their definition of synergy $\mathcal{S}_{\mathrm{VK}}$ must be invariant to duplicates for this distribution, $\Pi^{\mathrm{sx}}(T:\{1,2\}\{2,3\}) = \Pi^{\mathrm{sx}}(T:\{1,2\})$, since the mutual information

is invariant to duplicates, $I(T : S_1, S_2, S_3) = I(T :, S_1, S_2)$. Also, they proved that $\mathcal{S}_{\text{VK}}$ is invariant to duplicates in general [104].

For the shared exclusion measure $i_\cap^{\text{sx}}$, it is evident that the invariant property will hold since the shared information is indeed a mutual information and it is easy to see that $i_\cap^{\text{sx}}(t : s_1; s_2; s_3) = i_\cap^{\text{sx}}(t : s_1; s_2)$. In fact, we show below that all the PID terms are invariant to the duplication. That is, the unique information of S_2 is invariant and captured by $\Pi^{\text{sx}}(T : \{2\})$. Also, the unique information of S_1 is invariant but is captured by the atom $\Pi^{\text{sx}}(T : \{1\}\{3\})$ since it is shared information by S_1 and S_3 as S_3 is a copy of S_1. Finally, the synergistic information is invariant, however, it is captured by $\Pi^{\text{sx}}(T : \{1, 2\}\{2, 3\})$ since the coalitions (S_1, S_2) and (S_2, S_3) can equally specify the target. These claims are shown below by replacing s_3 by s_1 and applying the monotonicity axiom 2 on $i_\cap^{\text{sx}+}$ and $i_\cap^{\text{sx}-}$. Note that due to symmetry all the realizations have equal PID terms and the difference between the informative and misinformative is computed implicitly.

For any (t, s_1, s_2, s_3) with nonzero probability mass, we have

$$
\begin{aligned}
&i_\cap^{\text{sx}}(t : s_1; s_2; s_3) = i_\cap^{\text{sx}}(t : s_1; s_2) = i_\cap^{\text{sx}}(t : s_2; s_3) = -0.5849 \\
&i_\cap^{\text{sx}}(t : s_1; s_3) = i_\cap^{\text{sx}}(t : s_1) = i_\cap^{\text{sx}}(t : s_3) = 0
\end{aligned}
$$

implying that

$$
\begin{aligned}
&\pi^{\text{sx}}(t : \{1\}\{2\}) = \pi^{\text{sx}}(t : \{2\}\{3\}) = 0 \\
&\pi^{\text{sx}}(t : \{1\}\{3\}) = -\pi^{\text{sx}}(t : \{1\}\{2\}\{3\}) = 0.5849.
\end{aligned}
$$

But, $i_\cap^{\text{sx}}(t : s_2; s_1, s_3) = i_\cap^{\text{sx}}(t : s_2; s_3) = i_\cap^{\text{sx}}(t : s_1; s_2) = -0.5849$ meaning that

$$
\begin{aligned}
&\pi^{\text{sx}}(t : \{2\}\{1, 3\}) = 0 \\
&\pi^{\text{sx}}(t : \{2\}) = i^{\text{sx}}(t : s_2) - i^{\text{sx}}(t : s_2; s_1, s_3) = 0.5849.
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
&i_\cap^{\text{sx}}(t : s_1; s_2, s_3) = i_\cap^{\text{sx}}(t : s_1; s_1, s_2) = i_\cap^{\text{sx}}(t : s_1) = 0 \\
&i_\cap^{\text{sx}}(t : s_3; s_1, s_2) = i_\cap^{\text{sx}}(t : s_1; s_1, s_2) = i_\cap^{\text{sx}}(t : s_1) = 0 \\
&i_\cap^{\text{sx}}(t : s_1, s_2; s_1, s_3; s_2, s_3) = i_\cap^{\text{sx}}(t : s_1) = 0
\end{aligned}
$$

and so

$$
\begin{aligned}
&\pi^{\text{sx}}(t : \{1\}\{2, 3\}) = \pi^{\text{sx}}(t : \{3\}\{1, 2\}) = 0 \\
&\pi^{\text{sx}}(t : \{1, 2\}\{2, 3\}) = 0.415 \\
&\pi^{\text{sx}}(t : \{1, 2\}\{1, 3\}) = \pi^{\text{sx}}(t : \{1, 2\}\{2, 3\}) = 0.
\end{aligned}
$$

Finally, we have

$$i_\cap^{\text{sx}}(t : s_1, s_2, s_3) = i_\cap^{\text{sx}}(t : s_1, s_2) = i_\cap^{\text{sx}}(t : s_2, s_3) = 1$$
$$i_\cap^{\text{sx}}(t : s_1, s_3) = 0$$

and thus it easy to see that their corresponding atoms are zero.

3.6.5 Probability distribution 3-bit parity

Let S_1, S_2 and S_3 be independent, uniformly distributed random bits, and $T = \sum_{i=1}^{3} S_i \bmod 2$. This distribution is the 3-bit parity, where T indicates the parity of the total number of 1-bits in (S_1, S_2, S_3). Note that all possible realizations occur with probability $1/8$ and result in the same PPID as well as the average PID due to the symmetry of the variables. Table 3.4 shows the informative and misinformative component, and their difference for any realization. In addition, we illustrate in Figure 3.4 the results of $\pi^{\text{sx}}(t : \{1, 2\}\{3, 4\})$ for the 4-bit parity distribution.

3.7 Appendix

3.7.1 Lattice structure: supporting proofs and further details

We show how the redundancy lattice can be endowed by $i_\cap^{\text{sx}\,\pm}$ separately to obtain consistent PID terms $\pi_\pm^{\text{sx}}$. Subsequently, we show that $\pi_\pm^{\text{sx}}$ are nonnegative and thus the PID terms are meaningful.

Informative and misinformative lattices

We start by explaining the redundancy lattice proposed by Williams and Beer. Then, we explain in detail how to apply $i_\cap^{\text{sx}}$ to obtain a PID.

As explained in section 3.4, there is a one-to-one correspondence between the PID terms and the antichain combinations. Since $i_\cap^{\text{sx}}$ is defined locally, then for every realization the antichain combinations are associated to the source events. This way the PPID terms are computed and their average amount to the desired PID terms.

We use specific index sets and call them *antichains* to represent the antichain combination since antichain combinations are uniquely identified by the indices of

Tab. 3.4: 3-bit Parity Example. *Left:* the average *informative* partial information decomposition is evaluated. *Right:* the average *misinformative* partial information decomposition is evaluated. *Center:* the average partial information decomposition is evaluated.

Π^{sx}_{+}				Π^{sx}				Π^{sx}_{-}			
	$\{1,2,3\}$ 0.2451				$\{1,2,3\}$ 0.2451				$\{1,2,3\}$ 0		
$\{1,2\}$ 0.1699	$\{1,3\}$ 0.1699	$\{2,3\}$ 0.1699		$\{1,2\}$ 0.1699	$\{1,3\}$ 0.1699	$\{2,3\}$ 0.1699		$\{1,2\}$ 0	$\{1,3\}$ 0	$\{2,3\}$ 0	
$\{1,2\}\{1,3\}$ 0.0931	$\{1,2\}\{2,3\}$ 0.0931	$\{1,3\}\{2,3\}$ 0.0931		$\{1,2\}\{1,3\}$ 0.0931	$\{1,2\}\{2,3\}$ 0.0931	$\{1,3\}\{2,3\}$ 0.0931		$\{1,2\}\{1,3\}$ 0	$\{1,2\}\{2,3\}$ 0	$\{1,3\}\{2,3\}$ 0	
$\{1\}$ 0.3219	$\{2\}$ 0.3219	$\{3\}$ 0.3219	$\{1,2\}\{1,3\}\{2,3\}$ 0.0182	$\{1\}$ 0.3219	$\{2\}$ 0.3219	$\{3\}$ 0.3219	$\{1,2\}\{1,3\}\{2,3\}$ -0.2268	$\{1\}$ 0.3219	$\{2\}$ 0.3219	$\{3\}$ 0.3219	$\{1,2\}\{1,3\}\{2,3\}$ 0.2451
$\{1\}\{2,3\}$ 0.0406	$\{2\}\{1,3\}$ 0.0406	$\{3\}\{1,2\}$ 0.0406		$\{1\}\{2,3\}$ -0.1293	$\{2\}\{1,3\}$ -0.1293	$\{3\}\{1,2\}$ -0.1293		$\{1\}\{2,3\}$ 0.1699	$\{2\}\{1,3\}$ 0.1699	$\{3\}\{1,2\}$ 0.1699	
$\{1\}\{2\}$ 0.2224	$\{1\}\{3\}$ 0.2224	$\{2\}\{3\}$ 0.2224		$\{1\}\{2\}$ -0.1926	$\{1\}\{3\}$ -0.1926	$\{2\}\{3\}$ -0.1926		$\{1\}\{2\}$ 0.415	$\{1\}\{3\}$ 0.415	$\{2\}\{3\}$ 0.415	
	$\{1\}\{2\}\{3\}$ 0.1926				$\{1\}\{2\}\{3\}$ 0.1926				$\{1\}\{2\}\{3\}$ 0		

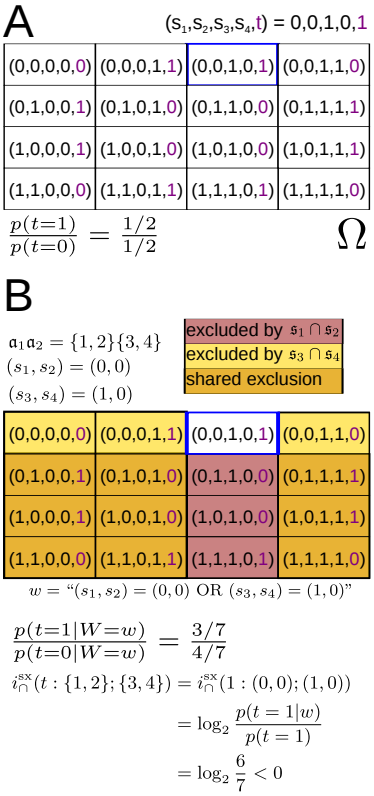

Fig. 3.4: Worked example of $i_\cap^{\mathrm{sx}}$ for a four source-variables case. We evaluate the shared information $i_\cap^{\mathrm{sx}}(t : \mathbf{a}_1; \mathbf{a}_2)$ with $\mathbf{a}_1 = \{1, 2\}$, $\mathbf{a}_2 = \{3, 4\}$, $s = (s_1, s_2, s_3, s_4) = (0, 0, 1, 0)$, and $t = \mathrm{Parity}(s) = 1$. (A) Sample space – the relevant event is marked by the blue (gray) outline. (B) exclusions induced by the two collections of source realization indices $\mathbf{a}_1$ (brown (dark gray)), $\mathbf{a}_2$ (yellow (light gray)), and the shared exclusion relevant for $i_\cap^{\mathrm{sx}}$ (gold (gray)). After removing and rescaling, the probability for the target event that was actually realized, i.e., $t = 1$, is reduced from $1/2$ to $3/7$. Hence the shared exclusion leads to negative shared information. Hence, $\pi^{\mathrm{sx}}(t : \{1, 2\}\{3, 4\}) = -0.0145\,\mathrm{bit}$.

their source events. For instance, an antichain $\alpha = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ such that $\mathbf{a}_i \subset [n]$ where $[n]$ is the index set of the realization $s = (s_1, \ldots, s_n)$. Moreover, $\mathbf{a}_i \in \alpha$ should be pairwise incomparable under inclusion since antichain combinations are as such (see Section 3.4). E.g., $\{\{1,2\},\{1,3\}\}$ represents the source event $(\mathfrak{s}_1 \cap \mathfrak{s}_2) \cup (\mathfrak{s}_1 \cap \mathfrak{s}_3)$ and the combination of (s_1, s_2) and (s_1, s_3).

Let $\mathscr{A}([n])$ be the set of all antichains; Crampton et al. [105] showed that there exists the following partial ordering over $\mathscr{A}([n])$:

$$\alpha \preceq \beta \Leftrightarrow \forall\ \mathbf{b} \in \beta, \exists\ \mathbf{a} \in \alpha \mid \mathbf{a} \subseteq \mathbf{b} \quad \forall\ \alpha, \beta \in \mathscr{A}([n]).$$

This partial ordering $\preceq$ implies that any $\alpha, \beta \in \mathscr{A}([n])$ have an infimum $\alpha \wedge \beta \in \mathscr{A}([n])$ and a supremum $\alpha \vee \beta \in \mathscr{A}([n])$ and so $\langle \mathscr{A}([n]), \preceq \rangle$ is called a lattice. Now when endowing $\langle \mathscr{A}([n]), \preceq \rangle$ with a function f (say a shared information) such that $f(\alpha) = \sum_{\beta \preceq \alpha} \pi(\beta)$ where $\pi(\beta)$ are desired quantities (say PID terms) that have a one-to-one correspondence with $\beta \in \mathscr{A}([n])$, then we can compute these π using f. Hence, we reduced the problem of defining different conceptual quantities that each antichain represents by defining a single conceptual quantity for each antichain that is the shared mutual information.

Williams and Beer coined this idea of endowing $\langle \mathscr{A}([n]), \preceq \rangle$ with a redundancy measure $I_\cap$ and hence the name "redundancy lattice." For this, they had a set of axioms that ensured (i) the one-to-one correspondence between $\mathscr{A}([n])$ and the PID terms and (ii) that $I_\cap(\alpha) = \sum_{\beta \preceq \alpha} \Pi(\beta)$. However, their definition was not local (for every realization) and thus Finn and Lizier [91] adapted the axioms for the local case. However, the local shared measure $i_\cap$ can take negative values and the problem persists upon averaging. Thus, they proposed to decompose $i_\cap = i_\cap^+ - i_\cap^-$ where $i_\cap^\pm$ take only nonnegative terms and can be interpreted as informative and misinformative components of $i_\cap$. Altogether, for each realization we will endow $\langle \mathscr{A}([n]), \preceq \rangle$ with $i_\cap^{\mathrm{sx}\,+}$ (*informative lattice*) and $i_\cap^{\mathrm{sx}\,-}$ (*misinformative lattice*) individually to obtain π_+^{sx} and π_-^{sx} PPID terms.

First, for any $\alpha \in \mathscr{A}[n]$, we define $i_\cap^{\mathrm{sx}\,\pm}$ as follows:

$$\begin{aligned}
\mathbb{P}(\alpha) &= \mathbb{P}\Big(\bigcup_{\mathbf{a}\in\alpha} \bigcap_{i\in\mathbf{a}} \mathfrak{s}_i\Big) \\
\mathbb{P}(\mathfrak{t}, \alpha) &= \mathbb{P}\Big(\bigcup_{\mathbf{a}\in\alpha} \bigcap_{i\in\mathbf{a}} (\mathfrak{t} \cap \mathfrak{s}_i)\Big) \\
i_\cap^{\mathrm{sx}}(t:\alpha) &= \log_2 \frac{1}{\mathbb{P}(\alpha)} - \log_2 \frac{\mathbb{P}(\mathfrak{t})}{\mathbb{P}(\mathfrak{t} \cap \alpha)} \\
&= i_\cap^{\mathrm{sx}\,+}(t:\alpha) - i_\cap^{\mathrm{sx}\,-}(t:\alpha).
\end{aligned}$$

Now to show that this endowing of $i_\cap^{\mathrm{sx}\pm}$ is consistent, we prove Theorem 4, that shows that $i_\cap^{\mathrm{sx}\pm}$ satisfy the PPID axioms.

proof of Theorem 4. By the symmetry of intersection, $i_\cap^{\mathrm{sx}\pm}$ defined in (3.14) satisfy the symmetry Axiom 1. For any collection $\mathbf{a}$, using (3.14), the informative and misinformative shared information are

$$i_\cap^{\mathrm{sx}+}(t:\mathbf{a}) = \log_2 \frac{1}{p(\mathbf{a})} = h(\mathbf{a})$$
$$i_\cap^{\mathrm{sx}-}(t:\mathbf{a}) = \log_2 \frac{p(t)}{p(t,\mathbf{a})} = h(\mathbf{a} \mid t).$$

and so they satisfy Axiom 3. For Axiom 2, note that

$$\mathbb{P}(\bar{\mathfrak{a}}_1, \bar{\mathfrak{a}}_2, \ldots, \bar{\mathfrak{a}}_m, \bar{\mathfrak{a}}_{m+1}) \leqslant \mathbb{P}(\bar{\mathfrak{a}}_1, \bar{\mathfrak{a}}_2, \ldots, \bar{\mathfrak{a}}_m)$$

This implies that $i_\cap^{\mathrm{sx}\pm}$ decrease monotonically if joint source realizations are added, where equality holds if there exists $i \in [m]$ such that $\bar{\mathfrak{a}}_{m+1} \supseteq \bar{\mathfrak{a}}_i$, i.e., if there exists $i \in [m]$ such that $\mathfrak{a}_{m+1} \subseteq \mathfrak{a}_i \Leftrightarrow \mathbf{a}_i \subseteq \mathbf{a}_{m+1}$. □

Then, we assume that

$$i_\cap^{\mathrm{sx}\pm}(t:\alpha) = \sum_{\beta \preceq \alpha} \pi_\pm^{\mathrm{sx}}(t:\beta) \quad \forall\, \alpha, \beta \in \mathscr{A}([n]). \tag{3.19}$$

Note that, this assumption is logically sound and is discussed thoroughly in [92]. Finally, to obtain $\pi_\pm^{\mathrm{sx}}$, we show that Eq. (3.19) is invertible via a so-called Möbius inversion given by the following theorem.

Theorem 7. *Let $i_\cap^{\mathrm{sx}\pm}$ be measures on the redundancy lattice, then we have the following closed form for each atom $\pi_\pm^{\mathrm{sx}}$:*

$$\pi_\pm^{\mathrm{sx}}(t:\alpha) = i_\cap^{\mathrm{sx}\pm}(t:\alpha) - \sum_{\varnothing \neq \mathcal{B} \subseteq \alpha^-} (-1)^{|\mathcal{B}|-1} i_\cap^{\mathrm{sx}\pm}(t : \bigwedge \mathcal{B}). \tag{3.20}$$

The proof of the above theorem follows from that of [91, Theorem A1].

Nonnegativity of $\pi_\pm^{\mathrm{sx}}$

In order for our information decomposition to be interpretative, the informative and misinformative atoms, $\pi_\pm^{\mathrm{sx}}$, must be nonnegative. First, we recall these results from convex analysis that will come in handy later.

Theorem 8 (Theorem 2.67 [106]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Then, f is convex if and only if for all x and y*

$$f(y) \geqslant f(x) + \nabla^T f(x)(y - x).$$

Proposition 5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable convex function and $y_0 - x_0 = c\mathbf{1}$ where $c \geqslant 0$. If $f(x_0) \geqslant f(y_0)$, then*

$$-\sum_i \frac{\partial f}{\partial x_i}(y_0) \leqslant -\sum_i \frac{\partial f}{\partial x_i}(x_0).$$

Proof. For any $x, y \in \mathbb{R}^n$, using theorem 8 by interchanging the roles of x and y,

$$-\nabla^T f(y)(y - x) \leqslant f(x) - f(y) \leqslant -\nabla^T f(x)(y - x).$$

Now consider $x_0, y_0 \in \mathbb{R}^n$ such that $y_0 - x_0 = c\mathbf{1}$, then

$$\begin{aligned} -c\nabla^T f(y_0)\mathbf{1} &\leqslant -c\nabla^T f(x_0)\mathbf{1} \\ -\sum_i \frac{\partial f}{\partial x_i}(y_0) &\leqslant -\sum_i \frac{\partial f}{\partial x_i}(x_0). \end{aligned}$$

□

We write down the proof of theorem 5 and then show that $i^{\text{sx}\pm}_{\cap}$ are nonnegative.

proof of theorem 5. Let $\alpha, \beta \in \mathscr{A}([n])$ and $\alpha \preceq \beta$. Then α and β are of the form $\alpha = \{\mathbf{a}_1, ..., \mathbf{a}_{k_\alpha}\}$ and $\beta = \{\mathbf{b}_1, ..., \mathbf{b}_{k_\beta}\}$. Because $\alpha \preceq \beta$ there is a function $f : \beta \to \alpha$ such that $f(\mathbf{b}) \subseteq \mathbf{b}$ [6]. Now we have for all $\mathbf{b} \in \beta$

$$\bigcap_{i \in \mathbf{b}} \mathfrak{s}_i \subseteq \bigcap_{i \in f(\mathbf{b})} \mathfrak{s}_i$$

Hence,

$$\begin{aligned} \mathbb{P}(\beta) = \mathbb{P}\left(\bigcup_{\mathbf{b}\in\beta} \bigcap_{i\in\mathbf{b}} \mathfrak{s}_i\right) &\leqslant \mathbb{P}\left(\bigcup_{\mathbf{b}\in\beta} \bigcap_{i\in f(\mathbf{b})} \mathfrak{s}_i\right) \\ &\leqslant \mathbb{P}\left(\bigcup_{\mathbf{a}\in\alpha} \bigcap_{i\in\mathbf{a}} \mathfrak{s}_i\right) = \mathbb{P}(\alpha). \end{aligned} \tag{3.21}$$

[6] This function does not have to be surjective: Suppose $\alpha = \{\{1\}, \{2, 4\}, \{3\}\}$ and $\beta = \{\{1, 2, 3, 4\}\}$. Then necessarily two sets in α will not be in the image of f. It also does not have to be injective. Consider $\alpha = \{1\}$ and $\beta = \{\{1, 2\}, \{1, 3\}\}$. Then both elements of β have to be mapped to the only element of α

The last inequality is true because the term on its L.H.S. is the probability of a union of intersections related to collections $\mathbf{a} \in \alpha$ (the $f(\mathbf{b})$), i.e., it is the probability of a union of events of the type $\bigcap_{i\in\mathbf{a}} \mathfrak{s}_i$. The probability of such a union can only get bigger if we take it over *all* events of this type. Using (3.21), it immediately follows that $i_\cap^{\mathrm{sx}+}(t:\alpha) \leqslant i_\cap^{\mathrm{sx}+}(t:\beta)$ and $i_\cap^{\mathrm{sx}+}$ is monotonically increasing. Using the same argument, $i_\cap^{\mathrm{sx}-}$ is monotonically increasing. □

Proposition 6. *$i_\cap^{\mathrm{sx}\pm}$ are nonnegative.*

Proof. $i_\cap^{\mathrm{sx}+}(t:\mathbf{a}_1;\mathbf{a}_2;\ldots;\mathbf{a}_m) = \log_2 \frac{1}{\mathbb{P}(\mathfrak{a}_1 \cup \mathfrak{a}_2 \cup \ldots \cup \mathbf{a}_m)} \geqslant 0.$

Similarly, the misinformative

$$i_\cap^{\mathrm{sx}-}(t:\mathbf{a}_1;\mathbf{a}_2;\ldots;\mathbf{a}_m) = \log_2 \frac{\mathbb{P}(\mathfrak{t})}{\mathbb{P}(\mathfrak{t} \cap [(\cap_{i\in\mathbf{a}_1} \mathfrak{s}_i) \cup (\cap_{i\in\mathbf{a}_2} \mathfrak{s}_i) \cup \ldots \cup (\cap_{i\in\mathbf{a}_m} \mathfrak{s}_i)])} \geqslant 0.$$

□

We construct a family of mappings from $\mathscr{P}(\alpha^-)$ where α^- is the set of children of α to the $\mathscr{A}([n])$ (see FIG 3.5). This family of mappings plays a key role in the desired proof of nonnegativity.

Proposition 7. *Let $\alpha \in \mathscr{A}([n])$ and $\alpha^- = \{\gamma_1, \ldots, \gamma_k\}$ ordered increasingly w.r.t. the probability mass be the set of children of α on $\langle \mathscr{A}([n]), \preceq \rangle$. Then, for any $1 \leqslant i \leqslant k$*

$$\begin{aligned} f_i : \mathscr{P}_1(\alpha^- \backslash \{\gamma_i\}) \cup \{\{\alpha\}\} &\longrightarrow \mathscr{A}([n]) \\ \mathcal{B} &\longrightarrow \bigwedge_{\beta\in\mathcal{B}} \beta \wedge \gamma_i \end{aligned}$$

is a mapping such that $\mathbb{P}(f_i(\mathcal{B})) = \mathbb{P}(\bigwedge_{\beta\in\mathcal{B}} \beta) + d_i$ where $d_i = \mathbb{P}(\gamma_i) - \mathbb{P}(\alpha)$ and the complement is taken w.r.t. $\mathscr{P}(\alpha^-)$, the powerset of α^-.

Proof. Since $\gamma_i \in \alpha^-$ and $\beta \in \alpha^-$ for any $\beta \in \mathcal{B}$, then $(\bigwedge_{\beta\in\mathcal{B}} \beta) \vee \gamma_i = \alpha$. Now, for any $\mathcal{B} \in \mathscr{P}(\alpha^- \backslash \{\gamma_i\})$, using the inclusion-exclusion, $\beta \wedge \gamma_i = \underline{\beta \cup \gamma_i}$ and $\beta \vee \gamma_i = \underline{\uparrow \beta \cap \uparrow \gamma_i}$,

$$\begin{aligned} \mathbb{P}(f_i(\mathcal{B})) = \mathbb{P}(\bigwedge_{\beta\in\mathcal{B}} \beta \wedge \gamma_i) &= \mathbb{P}(\bigwedge_{\beta\in\mathcal{B}} \beta) + \mathbb{P}(\gamma_i) - \mathbb{P}(\bigwedge_{\beta\in\mathcal{B}} \beta \vee \gamma_i) \\ &= \mathbb{P}(\beta) + \mathbb{P}(\gamma_i) - \mathbb{P}(\alpha). \end{aligned}$$

□

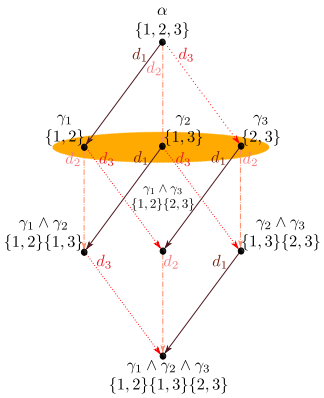

Fig. 3.5: The family of mappings introduced in proposition 7 that preserve the probability mass difference. Let α be the top node of $\mathscr{A}([3])$. The orange (gray dotted) region is α^-, the set of children of α. Each color depicts one mapping in the family based on some $\gamma \in \alpha^-$. The dark red (solid line) mapping is based on γ_1, the red mapping (dash-dotted line) is based on γ_2 and the salmon (dotted line) mapping is based on γ_3.

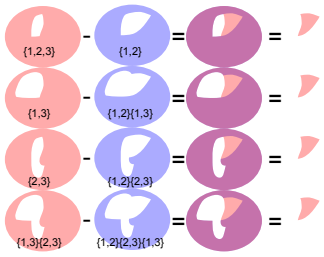

Fig. 3.6: Depiction of set differences corresponding to the probability mass difference d_1 introduced in proposition 7 and shown in Fig. 3.5, for the sets from Fig. 3.2.

The following lemma shows that for any node $\alpha \in \mathscr{A}([n])$, the recursive Eq. (3.20) should be nonnegative which is the main point in the desired proof of nonnegativity.

Lemma 2. *Let $\alpha \in \mathscr{A}([n])$; then*

$$-\log_2 \mathbb{P}(\alpha) + \sum_{\varnothing \neq \mathcal{B} \subseteq \alpha^-} (-1)^{|\mathcal{B}|-1} \log_2 \mathbb{P}(\bigwedge \mathcal{B}) \geqslant 0. \tag{3.22}$$

Proof. Suppose that $|\alpha^-| = k$ and w.l.o.g. that $\alpha^- = \{\gamma_1, \ldots, \gamma_k\}$ is ordered increasingly w.r.t. the probability mass. The proof will follow by induction over $k = |\alpha^-|$. We will demonstrate the inequality (3.22) for $k = 3, 4$ to show the induction basis. For $k = 3$, the L.H.S. of (3.22) can be written as

$$\begin{aligned}
&\log_2 \frac{\mathbb{P}(\gamma_1)\,\mathbb{P}(\gamma_2)\,\mathbb{P}(\gamma_3)\,\mathbb{P}(\gamma_1 \wedge \gamma_2 \wedge \gamma_3)}{\mathbb{P}(\alpha)\,\mathbb{P}(\gamma_1 \wedge \gamma_2)\,\mathbb{P}(\gamma_1 \wedge \gamma_3)\,\mathbb{P}(\gamma_2 \wedge \gamma_3)} \\
&= \log_2 \frac{\frac{\mathbb{P}(\alpha)+d_1}{\mathbb{P}(\alpha)}}{\frac{(\mathbb{P}(\alpha)+d_2)+d_1}{(\mathbb{P}(\alpha)+d_2)}} - \log_2 \frac{\frac{\mathbb{P}(\alpha)+d_3+d_1}{\mathbb{P}(\alpha)+d_3}}{\frac{(\mathbb{P}(\alpha)+d_3+d_2)+d_1}{(\mathbb{P}(\alpha)+d_3+d_2)}} \\
&= [h_3(\mathbb{P}(\alpha)) - h_3(\mathbb{P}(\alpha) + d_2)] \\
&- [h_3(\mathbb{P}(\alpha) + d_3) - h_3(\mathbb{P}(\alpha) + d_3 + d_2)],
\end{aligned}$$

where $h_3(x) = \log_2(1+{}^{d_1}\!/_x)$, $d_i := \mathbb{P}(\gamma_i) - \mathbb{P}(\alpha)$ for $i \in \{1, 2, 3\}$, and $d_3 \geqslant d_2 \geqslant d_1 \geqslant 0$. Note that h_3 is a continuously differentiable convex function that is monotonically decreasing. Now, take $x = \mathbb{P}(\alpha)$ and $y = \mathbb{P}(\alpha) + d_3$, then

$$\begin{aligned}
&h_3(\mathbb{P}(\alpha)) - h_3(\mathbb{P}(\alpha) + d_2) \\
&\overset{\text{Thm. } 8}{\geqslant} -d_2 h_3'(\mathbb{P}(\alpha) + d_2) \\
&\overset{\text{Prop. } 5}{\geqslant} -d_2 h_3'(\mathbb{P}(\alpha) + d_3) \\
&\overset{\text{Thm. } 8}{\geqslant} h_3(\mathbb{P}(\alpha) + d_3) - h_3(\mathbb{P}(\alpha) + d_3 + d_2)
\end{aligned}$$

and so the inequality (3.22) holds when $k = 3$. For $k = 4$, we have $\alpha^- = \{\gamma_1, \gamma_2, \gamma_3, \gamma_4\}$ ordered increasingly w.r.t. the probability mass. By Proposition 7, the L.H.S. of (3.22) can be written as

$$
\begin{aligned}
&\Big[h_3\big(\mathbb{P}(\alpha)\big) - h_3\big(\mathbb{P}(\alpha) + d_2\big) - \Big(h_3(\mathbb{P}(\alpha) + d_3) \\
&- h_3\big(\mathbb{P}(\alpha) + d_3 + d_2\big)\Big)\Big] \\
&- \Big[h_3\big(\mathbb{P}(\alpha) + d_4\big) - h_3\big(\mathbb{P}(\alpha) + d_4 + d_2\big) \\
&- \Big(h_3\big(\mathbb{P}(\alpha) + d_4 + d_3\big) - h_3\big(\mathbb{P}(\alpha) + d_4 + d_3 + d_2\big)\Big)\Big] \\
&= \Big[h_4\big(\mathbb{P}(\alpha), \mathbb{P}(\alpha) + d_2\big) - h_4\big(\mathbb{P}(\alpha) + d_3, \mathbb{P}(\alpha) + d_3 + d_2\big)\Big] \\
&- \Big[h_4\big(\mathbb{P}(\alpha) + d_4, \mathbb{P}(\alpha) + d_4 + d_2\big) \\
&- h_4\big(\mathbb{P}(\alpha) + d_4 + d_3, \mathbb{P}(\alpha) + d_4 + d_3 + d_2\big)\Big],
\end{aligned}
$$

where $d_i := \mathbb{P}(\gamma_i) - \mathbb{P}(\alpha)$ for $i \in \{2, 3, 4\}$, $d_4 \geqslant d_3 \geqslant d_2 \geqslant 0$, and $h_4(x_1, x_2) = \log_2(1 + {}^{d_1(x_2 - x_1)}/_{x_1(x_2 + d_1)}) = h_3(x_1) - h_3(x_2)$. Let $\delta \geqslant 0$ and $x, y \in H_4^\delta := \{x \in \mathbb{R}_+^{2*} \mid x_2 = x_1 + \delta\}$ where $x_1 \leqslant y_1$, then $h_4(x) \geqslant h_4(y)$ since (3.22) holds for $k = 3$. Moreover, h_4 is convex since for any $x, y \in H_4^\delta$ and $\theta \in [0, 1]$

$$
\begin{aligned}
&\theta h_4(x) + (1 - \theta)h_4(y) - h_4(\theta x + (1 - \theta)y) \\
&= \theta(h_3(x_1) - h_3(x_2)) + (1 - \theta)(h_3(y_1) \\
&- h_3(y_2)) - h_3(\theta x_1 + (1 - \theta)y_1) + h_3(\theta x_2 + (1 - \theta)y_2) \\
&= [\theta h_3(x_1) + (1 - \theta)h_3(y_1) - h_3(\theta x_1 + (1 - \theta)y_1)] \\
&- [\theta h_3(x_1 + \delta) + (1 - \theta)h_3(y_1 + \delta) \\
&- h_3(\theta x_1 + (1 - \theta)y_1 + \delta)] \geqslant 0.
\end{aligned}
$$

Now, take $x = (\mathbb{P}(\alpha), \mathbb{P}(\alpha) + d_2)$ and $y = (\mathbb{P}(\alpha) + d_4, \mathbb{P}(\alpha) + d_4 + d_2)$, then

$$
\begin{aligned}
&h_4(\mathbb{P}(\alpha), \mathbb{P}(\alpha) + d_2) - h_3(p(\alpha) + d_3, \mathbb{P}(\alpha) + d_3 + d_2) \\
&\overset{\text{Thm. } 8}{\geqslant} -\nabla^T h_4(\mathbb{P}(\alpha) + d_3, \mathbb{P}(\alpha) + d_3 + d_2)(d_3, d_3) \\
&\overset{\text{Prop. } 5}{\geqslant} -\nabla^T h_4(\mathbb{P}(\alpha) + d_4, \mathbb{P}(\alpha) + d_4 + d_2)(d_3, d_3) \\
&\overset{\text{Thm. } 8}{\geqslant} h_4(\mathbb{P}(\alpha) + d_4, \mathbb{P}(\alpha) + d_4 + d_2) \\
&- h_4(\mathbb{P}(\alpha) + d_4 + d_3, \mathbb{P}(\alpha) + d_4 + d_3 + d_2),
\end{aligned}
$$

and so the inequality (3.22) holds.

Suppose that the inequality holds for k and let us proof it for $k+1$. Here $\alpha^- = \{\gamma_1, \gamma_2, \ldots, \gamma_{k+1}\}$ and using Proposition 7, the L.H.S. of (3.22) can be written as

$$
\begin{aligned}
&\Big[h_k\big(a_{k-2}\big) - h_k\big(a_{k-2} + d_{k-1}\mathbf{1}_{k-2}\big) - \Big(h_k\big(a_{k-2} + d_k\mathbf{1}_{k-2}\big) \\
&- h_k\big(a_{k-2} + (d_k + d_{k-1})\mathbf{1}_{k-2}\big)\Big)\Big] \\
&- \Big[h_k\big(a_{k-2} + d_{k+1}\mathbf{1}_{k-2}\big) - h_k\big(a_{k-2} + (d_{k+1} + d_{k-1})\mathbf{1}_{k-2}\big) \\
&- \Big(h_k\big(a_{k-2} + (d_{k+1} + d_k)\mathbf{1}_{k-2}\big) \\
&- h_k\big(a_{k-2} + (d_{k+1} + d_k + d_{k-1})\mathbf{1}_{k-2}\big)\Big)\Big] \\
&= \Big[h_{k+1}\big(a_{k-2}, a_{k-2} + d_{k-1}\mathbf{1}_{k-2}\big) \\
&- \Big(h_{k+1}\big(a_{k-2} + d_k\mathbf{1}_{k-2}, a_{k-2} + (d_k + d_{k-1})\mathbf{1}_{k-2}\big)\Big)\Big] \\
&- \Big[h_{k+1}\big(a_{k-2} + d_{k+1}\mathbf{1}_{k-2}, a_{k-2} + (d_{k+1} + d_{k-1})\mathbf{1}_{k-2}\big) \\
&- h_{k+1}\big(a_{k-2} + (d_{k+1} + d_k)\mathbf{1}_{k-2}, a_{k-2} + (d_{k+1} + d_k \\
&+ d_{k-1})\mathbf{1}_{k-2}\big)\Big]
\end{aligned}
$$

where $a_{k-2} := (\mathbb{P}(\alpha), \ldots, \mathbb{P}(\alpha) + \sum_{i=2}^{k-2} d_i) \in \mathbb{R}^{2^{k-2}}$, $d_i := \mathbb{P}(\gamma_i) - \mathbb{P}(\alpha)$ for $i \in \{2, \ldots, k+1\}$, $d_{k+1} \geqslant \cdots \geqslant d_2 \geqslant 0$, and $h_{k+1}(x_1, \ldots, x_{2^{k-1}}) = h_k(x_1, \ldots, x_{2^{k-2}}) - h_k(x_{2^{k-2}+1}, \ldots, x_{2^{k-1}})$.

Let $\delta \geqslant 0$ and $x, y \in H_{k+1}^{\delta} := \{x \in \mathbb{R}^{2^{k-1}} \mid x_i = x_j + \delta, i = j \bmod 2^{k-2}\}$ where $x_i \leqslant y_i$ for all i, then $h_{k+1}(x) \geqslant h(y)$ because the Ineq. (3.22) holds for k. Moreover, h_{k+1} is convex since for any $x, y \in H_{k+1}^{\delta}$ and $\theta \in [0, 1]$

$$
\begin{aligned}
&\theta h_{k+1}(x_1, \ldots, x_{2^{k-1}}) + (1-\theta) h_{k+1}(y_1, \ldots, y_{2^{k-1}}) \\
&- h_{k+1}(\theta x_1 + (1-\theta) y_1, \ldots, \theta x_{2^{k-1}} + (1-\theta) y_{2^{k-1}}) \\
&= \Big[\theta h_k(x_1, \ldots, x_{2^{k-2}}) + (1-\theta) h_k(y_1, \ldots, y_{2^{k-2}}) \\
&- h_k(\theta x_1 + (1-\theta) y_1, \ldots, \theta x_{2^{k-1}} + (1-\theta) y_{2^{k-2}})\Big] - \Big[\\
&\theta h_k(x_1 + \delta, \ldots, x_{2^{k-2}} + \delta) + (1-\theta) h_k(y_1 + \delta, \ldots, y_{2^{k-2}} + \delta) \\
&- h_k(\theta x_1 + (1-\theta) y_1 + \delta, \ldots, \theta x_{2^{k-2}} + (1-\theta) y_{2^{k-2}} + \delta)\Big].
\end{aligned}
$$

is nonnegative. Now, take $x = (a_{k-2}, a_{k-2}+d_{k-1}\mathbf{1}_{k-2})$ and $y = (a_{k-2}+d_{k+1}\mathbf{1}_{k-2}, a_{k-2}+(d_{k+1}+d_{k-1})\mathbf{1}_{k-2})$, then

$$\begin{aligned}
&h_{k+1}(a_{k-2}, a_{k-2}+d_{k-1}\mathbf{1}_{k-2}) - h_{k+1}(a_{k-2}+d_k\mathbf{1}_{k-2},\\
&a_{k-2}+(d_k+d_{k-1})\mathbf{1}_{k-2})\\
&\geqslant -d_k\nabla^T h_{k+1}(a_{k-2}+d_k\mathbf{1}_{k-2}, a_{k-2}+(d_k+d_{k-1})\mathbf{1}_{k-2})\mathbf{1}_{k-1}\\
&\geqslant -d_k\nabla^T h_{k+1}(a_{k-2}+d_{k+1}\mathbf{1}_{k-2},\\
&a_{k-2}+(d_{k+1}+d_{k-1})\mathbf{1}_{k-2})\mathbf{1}_{k-1}\\
&\geqslant h_{k+1}(a_{k-2}+d_{k+1}\mathbf{1}_{k-2}, a_{k-2}+(d_{k+1}+d_{k-1})\mathbf{1}_{k-2})\\
&- h_{k+1}(a_{k-2}+(d_{k+1}+d_k)\mathbf{1}_{k-2},\\
&a_{k-2}+(d_{k+1}+d_k+d_{k-1})\mathbf{1}_{k-2}),
\end{aligned}$$

where the first and third inequalities hold using theorem 8 and the second inequality holds using Proposition 5 and so the inequality (3.22) holds for $k+1$.

□

Finally we write down the proof of theorem 6 to conclude that $i_\cap^{\mathrm{sx}}$ yields meaningful PPID terms.

proof of theorem 6. For any $\alpha \in \mathscr{A}([n])$,

$$\begin{aligned}
\pi_+^{\mathrm{sx}}(t:\alpha) &= i_\cap^{\mathrm{sx}+}(t:\alpha) - \sum_{\varnothing\neq\mathcal{B}\subseteq\alpha^-} (-1)^{|\mathcal{B}|-1} i_\cap^{\mathrm{sx}+}(t:\bigwedge\mathcal{B})\\
&= -\log_2\mathbb{P}(\alpha) + \sum_{\varnothing\neq\mathcal{B}\subseteq\alpha^-} (-1)^{|\mathcal{B}|-1}\log_2\mathbb{P}(\bigwedge\mathcal{B}).
\end{aligned}$$

So, by Lemma 2 $\pi_+^{\mathrm{sx}}(t:\alpha) \geqslant 0$. Similarly, $\pi_-^{\mathrm{sx}}(t:\alpha) \geqslant 0$ since intersecting with t has no effect on the nonnegativity shown in Lemma 2. □

3.7.2 Definition of $i_\cap^{\mathrm{sx}}$ starting from a general probability space

Let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a probability space and $S_1, ..., S_n, T$ be discrete and finite random variables on that space, i.e.,

$$\begin{aligned}
S_i &: \Omega \to \mathcal{A}_{S_i}, \quad (\mathfrak{A}, \mathscr{P}(\mathcal{A}_{S_i})) - \text{measurable}\\
T &: \Omega \to \mathcal{A}_T, \quad (\mathfrak{A}, \mathscr{P}(\mathcal{A}_T)) - \text{measurable},
\end{aligned}$$

where $\mathcal{A}_{S_i}$ and $\mathcal{A}_T$ are the finite alphabets of the corresponding random variables and $\mathscr{P}(\mathcal{A}_{S_i})$ and $\mathscr{P}(\mathcal{A}_T)$ are the power sets of these alphabets. Given a subset

of source realization indices $\mathbf{a} \subseteq \{1, ..., n\}$ the *local mutual information* of source realizations $(s_i)_{i \in \mathbf{a}}$ about the target realization t is defined as

$$i(t : (s_i)_{i \in \mathbf{a}}) = i(t : \mathbf{a}) = \log_2 \frac{\mathbb{P}\left(\mathfrak{t} | \bigcap_{i \in \mathbf{a}} \mathfrak{s}_i\right)}{\mathbb{P}(\mathfrak{t})}.$$

The *local shared information* of an antichain $\alpha = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ (representing a set of collections of source realizations) about the target realization $t \in \mathcal{A}_T$ is defined in terms of the original probability measure $\mathbb{P}$ as a function $i_\cap^{\mathrm{sx}} : \mathcal{A}_T \times \mathscr{A}(s) \to \mathbb{R}$ with

$$i_\cap^{\mathrm{sx}}(t : \alpha) = i_\cap^{\mathrm{sx}}(t : \mathbf{a}_1; \ldots; \mathbf{a}_m) := \log_2 \frac{\mathbb{P}\left(\mathfrak{t} | \bigcup_{i=1}^m \mathfrak{a}_i\right)}{\mathbb{P}(\mathfrak{t})}.$$

A special case of this quantity is the local shared information of a *complete* sequence of source realizations $(s_1, \ldots, s_n)$ about the target realization t. This is obtained by setting $\mathbf{a}_i = \{i\}$ and $m = n$:

$$i_\cap^{\mathrm{sx}}(t : \{1\}; \ldots; \{n\}) = \log_2 \frac{\mathbb{P}\left(\mathfrak{t} | \bigcup_{i=1}^n \mathfrak{s}_i\right)}{\mathbb{P}(\mathfrak{t})}.$$

In contrast to other shared information terms, this is an *atomic* quantity corresponding to the very bottom of the lattice of antichains. Rewriting $i_\cap^{\mathrm{sx}}$ allows us to decompose it into the difference of two positive parts:

$$\begin{aligned} i_\cap^{\mathrm{sx}}(t : \mathbf{a}_1, ..., \mathbf{a}_m) = \log_2 \frac{\mathbb{P}\left(\mathfrak{t} \cap \bigcup_{i=1}^m \mathfrak{a}_i\right)}{\mathbb{P}(\mathfrak{t})\mathbb{P}\left(\bigcup_{i=1}^m \mathfrak{a}_i\right)} &= \log_2 \frac{1}{\mathbb{P}\left(\bigcup_{i=1}^m \mathfrak{a}_i\right)} \\ - \log_2 \frac{\mathbb{P}(\mathfrak{t})}{\mathbb{P}\left(\mathfrak{t} \cap \bigcup_{i=1}^m \mathfrak{a}_i\right)}&, \end{aligned}$$

using standard rules for the logarithm. We call

$$i_\cap^{\mathrm{sx}+}(t : \mathbf{a}_1, \ldots, \mathbf{a}_m) := \log_2 \frac{1}{\mathbb{P}\left(\bigcup_{i=1}^m \mathfrak{a}_i\right)}$$

the *informative* local shared information and

$$i_\cap^{\mathrm{sx}-}(t : \mathbf{a}_1, \ldots, \mathbf{a}_m) := \log_2 \frac{\mathbb{P}(\mathfrak{t})}{\mathbb{P}\left(\mathfrak{t} \cap \bigcup_{i=1}^m \mathfrak{a}_i\right)}$$

the *misinformative* local shared information.

3.8 Acknowledgments

We would like to thank Nils Bertschinger, Joe Lizier, Conor Finn and Robin Ince for fruitful discussions on PID. We would also like to thank Patricia Wollstadt, Viola Priesemann, Raul Vicente, Johannes Zierenberg, Lucas Rudelt and Fabian Mikulasch for their valuable comments on this paper.

3.9 Author contributions

MW initially proposed the shared exclusion measure for redundant information. AM analytically derived its essential mathematical properties. AG spelled out the interpretation of the measure in terms of logical statements, provided the measure-theoretic formulation, and critically reviewed the mathematical aspects of the paper. All authors were actively involved in writing and revising the manuscript and have read and approved the final version for publication.

4

Bits and Pieces: Understanding Information Decomposition from Part-whole Relationships and Formal Logic

Aaron J. Gutknecht [1*] Michael Wibral [1], Abdullah Makkeh [1]

1 Campus Institute for Dynamics of Biological Networks, Georg-August University, Goettingen, Germany

Published as: ***Gutknecht, A. J., Wibral, M., & Makkeh, A. (2021). Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. Proceedings of the Royal Society A, 477(2251), 20210110.***

Abstract

Partial information decomposition (PID) seeks to decompose the multivariate mutual information that a set of source variables contains about a target variable into basic pieces, the so called "atoms of information". Each atom describes a distinct way in which the sources may contain information about the target. For instance, some information may be contained uniquely in a particular source, some information may be shared by multiple sources, and some information may only become accessible synergistically if multiple sources are combined. In this paper we show that the entire theory of PID can be derived, firstly, from considerations of part-whole relationships between information atoms and mutual information terms, and secondly, based on a hierarchy of logical constraints describing how a given information atom can be accessed. In this way, the idea of a partial information decomposition is developed on the basis of two of the most elementary relationships in nature: the part-whole relationship and the relation of logical implication. This unifying

perspective provides insights into pressing questions in the field such as the possibility of constructing a PID based on concepts other than redundant information in the general n-sources case. Additionally, it admits of a particularly accessible exposition of PID theory.

4.1 Introduction

Partial information decomposition (PID) is an example of a rare class of problems where a deceptively simple question has perplexed researchers for many years, leading to heated disputes over possible solutions [107], simple but incomplete answers [3], and even to statements that the question should not be asked [108]. The core question of PID is how the information carried by multiple source variables about a target variable is distributed over the source variables. In other words, it is the information theoretic question of 'who knows what about the target variable'. Intuitively, answering this question involves finding out which information we could get from multiple variables alike (called redundant or shared information), which information we could get only from specific variables, but not the others (called unique information), and which information we can only obtain when looking at some variables together (called synergistic information).

Examples of questions involving PID, are found in almost all fields of quantitative research. In neuroscience, for instance, we are interested in how the activity of multiple neurons, that were recorded in response to a stimulus, can provide information about (i.e. encode) the stimulus. Specifically, we are interested in whether the information provided by those neurons about the stimulus is provided redundantly, such that we can obtain it from many (or any) of the recorded neural responses, or whether certain aspects are only present uniquely in individual neurons, but not others; finally, we may find that we need to analyze all neural responses together to decode the stimulus - a case of synergy. All three ways of providing information about the stimulus may coexist and the aim of PID analysis is to determine to what degree each of them is present [89].

In this way PID can be used as a framework for systematically testing and comparing theories of neural processing (such as predictive coding [109] or coherent infomax [88]) in terms of their information theoretic "footprint", i.e. in terms of the amounts of unique, redundant or synergistic information processing predicted by the theory. The key idea is to identify such theories with a specific information theoretic goal function (e.g. "maximize redundancy while at the same time allowing for a certain

degree of unique information"). One may then investigate empirically whether a given neural circuit in fact maximizes the goal function in question or one may use the PID framework to come up with entirely new goal functions [31].

The PID problem also arises in cryptography in the context of so called "secret sharing" [110]. The idea is that a multiple participants (the sources) each hold some partial information about a particular piece of information called the secret (the target). However, the secret can only be accessed if certain participants combine their information. In this context, PID describes how access to the secret is distributed over the participants.

The partial information decomposition framework has furthermore been used to to operationalize several core concepts in the study of complex and computational systems. These concepts include for instance the notion of information modification [101, 102] which has been suggested along with information storage and transfer as one of three fundamental component processes of distributed computation. It has also been proposed that the concepts of emergence and self-organisation can be made quantifiable within the PID framework [111],[26].

Despite the universality of the PID problem, solutions have only arisen very recently, and the work on consolidating and on distilling them into a coherent structure is still in progress. In this paper we aim to do so by rederiving the theory of partial information decomposition from the perspective of mereology (the study of parthood relations) and formal logic. The general structure of PID arrived at in this way is equivalent to the one originally described by Williams and Beer [19]. However, our derivation has the advantage of tackling the problem directly from the perspective of the *parts* into which the information carried by the sources about the target is decomposed, the so called "atoms of information". By contrast, the formulation used until now takes an indirect approach via the concept of redundant information. Furthermore, the approach described here is based on particularly elementary concepts: parthood between information contributions and logical implication between statements about source realizations.

The remainder of this paper is structured as follows: First, in §4.2 we derive the general structure underlying partial information decomposition from considerations of elementary parthood relationships between information contributions. This structure is general in the sense that it still leaves open the possibility for multiple alternative measures of information decomposition. We show that the axioms underlying the formulation by Williams and Beer [19, 91] can be proven within the framework described here. In §4.3 we utilize formal logic to derive a specific PID measure and in this way provide a complete solution to the information decomposition problem.

§4.4 shows that there is an intriguing connection between formal logic and PID in that the mathematical lattice structure underlying information decomposition is isomorphic to a lattice of logical statements ordered by logical implication. This gives rise to a completely independent exposition of PID theory in terms of a hierarchy of logical constraints on how information about the target can be accessed. In §4.5 we show that the ideas presented here can be utilized to systematically answer the question of whether a (full n-sources) PID can be induced by measures other than redundant information such as synergy or unique information. Before concluding in §4.7, we briefly address the important distinction between parthood relations and quantitative relations in §4.6.

4.2 The parthood perspective

Suppose there are n source variables $S_1, \ldots, S_n$ carrying some joint mutual information $I(T : S_1, \ldots, S_n)$ [1, 2] about some target variable T (see Figure 4.1, left). The goal of partial information decomposition is to decompose this joint mutual information into its component parts, the so called *atoms* of information. As explained in the introduction, these parts are supposed to represent unique, redundant, and synergistic information contributions. Now, what distinguishes these contributions are their defining part-whole relationships to the information provided by the different source variables: the information uniquely associated with one of the sources in only part of the information provided by *that* source and not part of the information provided by any other source. The information provided redundantly by multiple sources is part of the information carried by *each* of these sources. And the information provided synergistically by multiple source is only part of the information carried by them jointly but not part of the information carried by any of them individually. For this reason, it seems natural to make the part-whole relationship between pieces of information the basic concept of PID. The goal of this section is to make this idea precise, and in this way, to open up a new perspective for thinking about partial information decomposition.

The underlying idea is that any theory should be put on the foundation of as simple and elementary concepts as possible. The part-whole relation is one of the most basic relationships in nature. It appears on all spatial and temporal scales: atoms are parts of molecules, planets are parts of solar systems, the phase of hyperpolarisation is part of an action potential, infancy is part of a human beings life. Moreover, it is not a purely scientific concept but is also ubiquitous in ordinary life: we say for instance, that a prime minister is part of the government or that a slice of pizza is

part of the whole pizza. This ubiquity makes it particularly easy to think in terms of part-whole relationships. We hope, therefore, that starting from this vantage point will provide a particularly accessible and intuitive exposition of partial information decomposition. This factor is of particular importance when it comes to the practical application of PID to specific scientific questions and the interpretation of the results of a PID analysis.

Developing the theory of partial information decomposition means that we have to answer three questions:

1. What are atoms of the decomposition supposed to mean, i.e. what *type* of information should they represent?
2. How many atoms are there for a given number of information sources?
3. How large are the different atoms of information given a specific joint probability distribution of sources and target? How many *bits* of information does each atom provide?

In the following sections we will tackle each of these questions in turn.

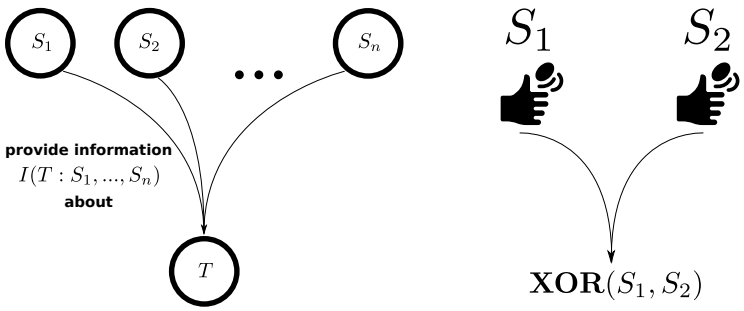

Fig. 4.1: Left: The general partial information decomposition problem is to decompose the joint mutual information provided by n source variables $S_1, \ldots, S_n$ about a target variables T into its component parts. Right: Illustration of the exclusive-or example. The sources are two independent coin flips. The target is 0 just in case both coins come up heads or both come up tails. It is 1 if one of the coins is heads while the other is tails. Coin tossing icons made by Freepik, www.flaticon.com.

4.2.1 What do the atoms of information mean?

Asking how to decompose the joint mutual information into its components parts is a bit like asking "How to slice a cake?". Of course, there are many possible ways to do so, and hence, there is no unique answer to the question. In order to make the question more precise we first have to provide a criterion according to which we would like to decompose the joint mutual information. This is what this section is

about. What are the atoms of information supposed to mean in the end, i.e. what *type* of information do they represent?

To a first approximation, the core idea underlying the parthood approach to partial information decomposition is to decompose the joint mutual information $I(T : S_1, \ldots, S_n)$ into information atoms, such that each atom is characterized by its parthood relations to the mutual information provided by the different sources. For instance, one atom of information will describe that part of the joint mutual information which is part of the information provided by *each* source, i.e. the information that is redundant to all sources. Another atom will describe the part of the joint mutual information that is only part of the information provided by the first source, i.e. it is unique to the first source. And so on.

Now, we have to refine this idea a bit: it is important to realize that it would not be enough to consider parthood relations to information provided by *individual* sources. The reason is that a *collection* of sources may provide some information that is not contained in any individual source but which only arises by *combining* the information from multiple sources in that collection. The classical example for this phenomenon is the logical exclusive-or shown in Figure 4.1, right. In this example the sources are two independent coin flips. The target is the exclusive-or of the sources, i.e. the target is 0 just in case both coins come up heads or both come up tails, and it is 1 otherwise. Initially, the odds for the target being zero or one respectively are 1:1 because there are four equally likely outcomes in two of which the target is 1 while it is 0 in the other two. Now, if we are told the value of one of the coins, these odds are not affected, and accordingly, we do not obtain any information about the target. For instance, if we are told that the first coin came up heads there are two equally likely outcomes left: Heads-Heads and Heads-Tails. In the first case, the target is zero and in the second case it is one. Hence, the odds are still 1:1. On the other hand, if we are told the value of *both* coins, then we *know* what the value of the target is. In other words, we obtain complete information about the target.

There are two conclusions to be drawn from examples like this:

1. There are cases in which multiple information sources combined provide some information that is not contained in any individual source. This type of information is generally called *synergistic information*.

2. Any reasonable theory of information should be compatible with the existence of synergistic information. In particular, it should allow that, in some cases,

the information provided jointly by multiple sources is larger than the sum of the individual information contributions provided by the sources.

Regarding the second point we may note that classical information theory satisfies this constraint because in some cases

$$I(T : S_1, S_2) > I(T : S_1) + I(T : S_2) \tag{4.1}$$

In fact, in the exclusive-or example, each individual source provides zero bits of information while the sources combined provide one bit of information.

Based on these consideration we may rephrase the basic idea of the parthood approach as: we are looking for a decomposition of the joint mutual information into atoms such that each atom is characterized by its parthood relations to the information carried by the different possible *collections* of sources about the target. Of course, we allow collections containing only a single source, such as $\{1\}$, as a special case. Note that we will generally refer to source variables and collections thereof *by their indices*. So instead of writing $\{S_1\}$ and $\{S_1, S_2\}$ to refer to the first source and the collection containing the first and second source, we write $\{1\}$ and $\{1, 2\}$ respectively. There are several important technical reasons for this that will become apparent in the following sections. For now it is sufficient to just think of it as a shorthand notation.

Let's now investigate how the idea of characterizing the information atoms by parthood relations plays out in the simple case of two sources S_1 and S_2. In this case, there are four collections:

1. The empty collection of sources $\{\}$
2. The collection containing only the first source $\{1\}$
3. The collection containing only the second source $\{2\}$
4. The collection containing both sources $\{1, 2\}$

Now, in order to characterize an information atom Π we have to ask for each collection $\mathbf{a}$: Is Π part of the information provided by $\mathbf{a}$? For two of the collections we can answer this question immediately for all Π: First, no atom of information should be contained in information provided by the empty collection of sources because there is no information in the empty set. If we do not know any source, then we cannot obtain any information from the sources. Second, any atom of information should be contained in the mutual information provided by the full set of sources since this is precisely what we want to decompose into its component

parts. Regarding the collections $\{1\}$ and $\{2\}$ we are free to answer yes or no leaving four possibilities as shown in Table 4.1.

Part of	{}	{1}	{2}	{1,2}
Π_1 (Synergy)	**0**	0	0	**1**
Π_2 (Unique)	**0**	1	0	**1**
Π_3 (Unique)	**0**	0	1	**1**
Π_4 (Shared)	**0**	1	1	**1**

Tab. 4.1: Parthood table for the case of two information sources. Each row characterizes a particular atom of information in terms of its parthood relationships with the mutual information provided by the different collections of sources. The bold entries are enforced by the constraints that there is no information in the empty collection of sources and that any piece of information is part of the information carried by the full set of sources about the target.

The first possibility (first row of Table 4.1) is an atom of information that is only part of the information provided by the sources jointly but not part of the information in either of the individual sources. This is the *synergistic information*. The second possibility (second row) is an atom that is part of the information provided by the first source but which is not part of the information in the second source. This atom of information describes the *unique information* of the first source. Similarly, the third possibility (third row) is an atom describing information uniquely contained in the second source. The fourth and last possibility (fourth row) is an atom that is part of the information provided by *each* source. This is the information *redundantly provided* or *shared by* the two sources.

So based on considerations of parthood we arrived at the conclusion that there should be exactly four atoms of information in the case of two source variables. Each atom is characterized by its parthood relations to the mutual information provided by the different collections of sources. These relationships are described by the rows of Table 4.1 which we will call *parthood distributions*. Each atom Π is formally represented by its parthood distribution f_Π.

Mathematically, a parthood distribution is a Boolean function from the powerset of $\{1, \ldots, n\}$ to $\{0, 1\}$, i.e. it takes a collection of source indices as an input and returns either 0 (the atom described by the distribution is not part of information provided by the collection) or 1 (the atom described by the distribution is part of that information) as an output. But note that not all such functions qualify as a parthood distribution. We already saw that certain constraints have to be satisfied. For instance, the empty set of sources has to be mapped to 0. We propose that there are exactly three constraints a parthood distribution f has to satisfy leading to the following definition

Definition 1. *A parthood distribution is any function* $f : \mathcal{P}(\{1, \ldots, n\}) \to \{0, 1\}$ *such that*

1. $f(\{\}) = 0$ *("There is no information in the empty set")*
2. $f(\{1, \ldots, n\}) = 1$ *("All information is in the full set")*
3. *For any two collections of source indices* $\mathbf{a}$, $\mathbf{b}$*: If* $\mathbf{b} \supseteq \mathbf{a}$*, then* $f(\mathbf{a}) = 1 \Rightarrow f(\mathbf{b}) = 1$ *(Monotonicity)*

The third constraint says that if an atom of information is part of the information provided by some collection of sources $\mathbf{a}$, then it also has to be part of the information provided by any superset of this collection. For example, if an atom is part of the information in source 1, then it also has to be part of the information in sources 1 and 2 combined. Note that this monotonicity constraint only matters if there are more than two information sources. Otherwise it is implied by the first two constraints. To fix ideas, an example of a Boolean function that is *not* a parthood distribution is shown in Table 4.2. The function assigns a 1 to the collection $\{1\}$ but a 0 to collections $\{1, 2\}$ and $\{1, 3\}$ which are supercollections of $\{1\}$. Thus, there can be no atom of information with the parthood relations described by this Boolean function.

Part of	{}	{1}	{2}	{3}	{1,2}	{1,3}	{2,3}	{1,2,3}
	0	1	0	0	**0**	**0**	0	1

Tab. 4.2: Example of Boolean function that is not a parthood distribution. Bold entries violate the monotonicity constraint.

We may now answer the question about the meaning of the atoms of information, i.e. what *type* of information they represent: They represent information that is part of the information provided by certain collections of sources but not part of the information of other collections. More precisely we can phrase this idea in terms of the following core principle:

Core Principle 1. *Each atom of information is characterized by a parthood distribution describing whether or not it is part of the information provided by the different possible collections of sources. The atom* $\Pi(f)$ *with parthood distribution f is exactly that part of the joint mutual information about the target which is 1) part of the information provided by all collections of sources* $\mathbf{a}$ *for which* $f(\mathbf{a}) = 1$*, and 2), which is not part of the information provided by collections for which* $f(\mathbf{a}) = 0$*.*

Given this characterization of the information atoms we are now in a position to answer the second question: How many atoms are there for a given number of information sources.

4.2.2 How many atoms of information are there?

Since each atom is characterized by its parthood distribution, the answer is straightforward: there is one atom per parthood distribution, or in other words, one atom per Boolean function satisfying the constraints presented in the previous section. The monotonicity constraint turns out to be most restrictive. In fact, once the monotonicity constraint is satisfied the other two constraints only rule out one Boolean function each as shown in Table 4.3. The reason is the following: Firstly, there is only a single *monotonic* Boolean function that assigns the value 1 to the empty set, namely, the function that is always 1. Since the empty set is subset of any other set, monotonicity enforces to assign a 1 to all sets once the empty set has value 1. However, this possibility is ruled out by the first constraint saying that there is no information in the empty set. Secondly, there is only a single *monotonic* Boolean function assigning the value 0 to the full set $\{1, \ldots, n\}$, namely the function that is always 0. Since any other set of source indices is contained in the full set, monotonicity forces us to assign a 0 to all sets once the full set has value 0. If we were to assign a 1 to any other set, then we would have to assign a 1 to the full set as well.

Part of	{}	...	...	...	{1,...,n}
	1 **0**	**1** **0**	**1** **0**	**1** **0**	**1** 0

Tab. 4.3: The two constant Boolean functions are ruled out by the first and second constraint on parthood distributions described above.

This means that the number of atoms is equal to *the number of monotonic Boolean functions minus two*. Now the sequence of the numbers of monotonic Boolean functions of n-bits is a very famous sequence in combinatorics called the *Dedekind numbers*. The Dedekind numbers are a very rapidly (in fact super-exponentially) growing sequence of numbers of which only the first eight entries are known to date [112]. The values for $2 \leqslant n \leqslant 6$ of the Dedekind numbers are: 6, 20, 168, 7581, 7828354.

Now that we have answered what type of information the different atoms represent and how many there are for a given number of information sources, there is one

important question left: How large are these different atoms? How many *bits* of information does each atom provide?

4.2.3 How large are the atoms of information?

The question of the sizes of the atoms is not a trivial one since the number of atoms grows so quickly. In the case of four information sources there are already 166 atoms. Hence, it does not appear to be feasible to define the amount of information of each of these atoms separately. What we need is a systematic approach that somehow fixes the sizes of all atoms at the same time. The core idea is to transform the problem into a much simpler one in which only a single type of informational quantity has to be defined. In the following we show how this can be achieved in three steps.

Define a quantitative relationship between atoms and composite quantities

So far we have only discussed how the atoms of information relate *qualitatively* to composite information quantities that are made up of multiple atoms, in particular mutual information (in the next section we will encounter another non-atomic quantity). We saw for instance, that in the case of two sources, the mutual information contributions provided by the individual sources, $I(T : S_1)$ and $I(T : S_2)$, each consist of a unique and a redundant information atom, while the joint mutual information $I(T : S_1, S_2)$ additionally consists of a synergistic part. This is illustrated in the information diagram shown in Figure 4.2.

Now the question arises: How are these mutual information terms related to the atoms they consist of *quantitatively*? The most straightforward answer (and the one generally accepted in the PID field) is that the mutual information is simply the *sum* of the atoms it consists of. We propose to extend this principle to any composite information quantity, i.e. any quantity that can be described as being made up out of multiple information atoms:

Core Principle 2. *The size of any non-atomic information quantity (i.e. the amount of information it contains) is the sum of the sizes of the information atoms it consists of.*

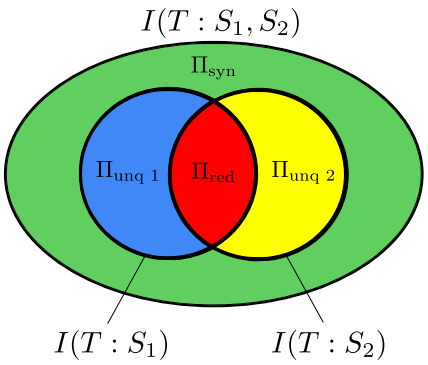

Fig. 4.2: Information diagram depicting the partial information decomposition for the case of two information sources. The inner two black circles represent the mutual information provided by the first source (left) and the second source (right) about the target. Each of these mutual information terms contains two atomic parts: $I(T : S_1)$ consists of the unique information in source 1 ($\Pi_{\text{unq }1}$, blue patch) and the information shared with source 2 (Π_{red}, red patch). $I(T : S_2)$ consists of the unique information in source 2 ($\Pi_{\text{unq }2}$, yellow patch) and again the shared information. The joint mutual information $I(T : S_1, S_2)$ is depicted by the large black oval encompassing the inner two circles. $I(T : S_1, S_2)$ consists of four atoms: The unique information in source 1 ($\Pi_{\text{unq }1}$, blue patch), the unique information in source 2 ($\Pi_{\text{unq }2}$, yellow patch), the shared information (Π_{red}, red patch), and additionally the synergistic information (Π_{syn}, green patch).

We could also rephrase this as "wholes are the sums of their (atomic) parts". In the case of two information sources, this principle leads to the following three equations:

$$I(T : S_1, S_2) = \Pi_{\text{red}} + \Pi_{\text{unq }1} + \Pi_{\text{unq }2} + \Pi_{\text{syn}} \tag{4.2}$$

$$I(T : S_1) = \Pi_{\text{red}} + \Pi_{\text{unq }1} \tag{4.3}$$

$$I(T : S_2) = \Pi_{\text{red}} + \Pi_{\text{unq }2} \tag{4.4}$$

This already gets us quite far in terms of determining the sizes of the atoms: The sizes of the atoms are the solutions to a linear system of equations. The only problem is that the system is underdetermined. We have four unknowns but only three equations. In the case of three sources, the problem is even more severe. In this case, there are seven non-empty collections of sources, and hence, seven mutual information terms. Again each of these terms is the sum of certain atoms. But as shown in section §4.2b there are 18 atoms. So we are short of 11 equations!

In general the equations relating the mutual information provided by some collection of sources $\mathbf{a}$ and the information atoms can be expressed easily in terms of their parthood distributions:

$$I(T : \mathbf{a}) = \sum_{f(\mathbf{a})=1} \Pi(f) \tag{4.5}$$

where $\Pi(f)$ is the information atom corresponding to parthood distribution f and the summation notation means that we are summing over all f such that $f(\mathbf{a}) = 1$. Note that on the left-hand-side we are using the shorthand notation $I(T : \mathbf{a})$ for the mutual information $I(T : (S_i)_{i \in \mathbf{a}})$ provided by the collection $\mathbf{a}$. Equation (4.5) can be taken to define a minimal notion of a partial information decomposition, i.e. any set of quantities $\Pi(f)$ at least has to satisfy this equation in order to be considered a partial information decomposition (or at least to be considered a parthood-based / Williams and Beer type PID). For a formal definition of such a minimally consistent PID see Appendix 4.8.1.

This concludes the first step. The next one is to find a way to come up with the appropriate number of additional equations. In doing so we will follow the same approach as Williams and Beer and utilize the concept of *redundant information* to introduce additional constraints. It should be noted that this is not the only way to derive a solution for the information atoms. In other words, a PID does not have to be "redundancy based". This issue is discussed in detail in §4.5. For now, however, let us follow the conventional path and see how it enables us to determine the sizes of the atoms of information.

Formulate additional equations using the concept of redundant information

The basic idea is now to extent the considerations of the previous step to another composite information quantity: the redundant information provided by multiple collections of sources about the target which we will generically denote by $I_\cap(T : \mathbf{a}_1, \ldots, \mathbf{a}_m)$. The $\cap$-symbol refers to the idea that the redundant information of collections $\mathbf{a}_1, \ldots, \mathbf{a}_m$ is the information contained in $\mathbf{a}_1$ *and* $\mathbf{a}_2$ *and*, ..., *and* $\mathbf{a}_m$. Intuitively, given two collections of sources $\mathbf{a}_1$ and $\mathbf{a}_2$, their redundant information is the information “shared” by those collections, what they have "in common", or geometrically: their overlap. These informal ideas are illustrated on the left side in Figure 4.3.

Note that the redundant information of multiple collections of information sources is not defined in classical information theory. We have to come up with an appropriate measure of redundant information ourselves. However, the informal ideas just describes already tell us that redundant information, no matter how we define it, should be related qualitatively to the information atoms in a very specific way: the information redundantly provided by multiple collections of sources should consist of exactly those information atoms that are part of the information carried by *all* of those collections:

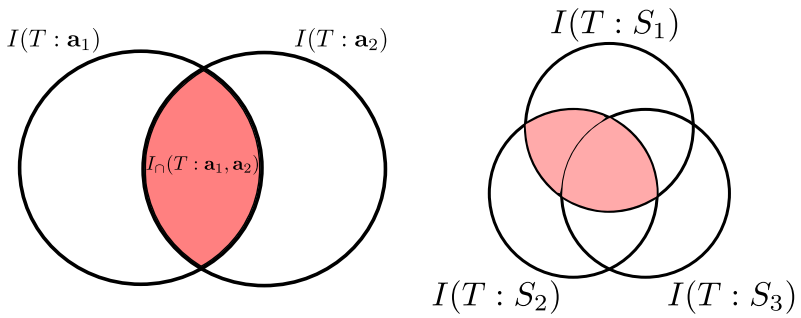

Fig. 4.3: Left: Illustration of the idea of the redundant information of collections $\mathbf{a}_1$ and $\mathbf{a}_2$. Right: Redundant information is generally not an atomic quantity. In the context of three information sources, the redundant information of sources 1 and 2 consists of two parts: the information shared by *only* by sources 1 and 2, and the information shared by all three sources.

Core Principle 3. *The redundant information* $I_\cap(T : \mathbf{a}_1, \ldots, \mathbf{a}_m)$ *consists of all information atoms that are part of the information provided by each* $\mathbf{a}_i$*, i.e. all atoms with a parthood distribution satisfying* $f(\mathbf{a}_i) = 1$ *for all* $i = 1, \ldots, m$*.*

Let's see what this principle implies in concrete examples. We saw that in the case of two sources, the redundant information of source 1 and source 2, $I(T : \{1\}, \{2\})$, is actually itself an atom, namely the atom with the parthood distribution

{}	{1}	{2}	{1,2}
0	1	1	1

This is the only atom that is part of both the information provided by the first source and also part of the information provided by the second source. But this is really a special case. Note what happens if we add a third source to the scenario. In this case the redundant information $I(T : \{1\}, \{2\})$ of sources 1 and 2 should consist of *two* parts: First, the information shared by *all three* sources (which is certainly also shared by sources 1 and 2), and secondly, the information shared *only* by sources 1 and 2 but not by source 3. This is illustrated on the right side in Figure 4.3. Note also that in the case of three sources there are actually *many* redundancies that we may compute:

1. the redundancy of all three sources $I_\cap(T : \{1\}, \{2\}, \{3\})$.
2. the redundancy of any *pair* of sources such as the redundancy of $I_\cap(T : \{1\}, \{2\})$.
3. the redundancy between a single source and a pair of sources such as $I_\cap(T : \{1\}, \{2, 3\})$.

4. the redundancy between two pairs of sources such as $I_\cap(T : \{1,2\},\{2,3\})$.

5. the redundancy of all three possible pairs of sources $I_\cap(T : \{1,2\},\{1,3\},\{2,3\})$.

It turns out that in total there are 11 redundancies (strictly speaking we should say 11 "proper" redundancies as will be explained below). But this is exactly the number of missing equations in the case of three information sources (see last paragraph of previous section).

Now, combining Core Principles 2 and 3, allows us the answer what the *quantitative* relationship between redundant information and information atoms has to be: the redundant information of collections of sources $\mathbf{a}_1,\ldots,\mathbf{a}_m$ is the sum of all atoms that are part of the information provided by *each* collection:

$$I_\cap(T : \mathbf{a}_1,\ldots,\mathbf{a}_m) = \sum_{f(\mathbf{a}_i)=1 \forall i=1,\ldots,m} \Pi(f) \tag{4.6}$$

where again the notation means that we are summing over all f that satisfy the condition below the summation sign. This equation can be read in two ways: First, as placing a constraint on the redundant information $I_\cap$, namely that it has to be the sum of specific atoms. This means that if we already knew the sizes of the Π's, we could compute $I_\cap$. However, the sizes of the Π's are precisely what we are trying to work out. Now the crucial idea is that we can also read the equation the other way around: if we can come up with some reasonable measure of redundant information $I_\cap$ we may be able to *invert* equation 4.6 in order to obtain the Π's. So the final step will be to show that such an inversion is in fact possible and will lead to a unique solution for the atoms of information.

Before proceeding to this step, it is important to briefly clarify the relationships between the three central concepts we have discussed so far:

1. the mutual information (the quantity we want to decompose)

2. the information atoms (the quantities we are looking for)

3. redundant information (the quantity we are going to use to find the information atoms)

These concept are easily confused with each other but should be clearly separated. The relationships between them are shown in Figure 4.4. First, based on what we have said so far, mutual information can be shown to be a special case of redundant information: the redundant information of a single collection $I_\cap(T : \mathbf{a}_1)$, i.e. "the information the collection shares *with itself* about the target". The reason

for this is that Core Principle 3 tells us that the redundant information of a single collection consists of all the atoms that are part of the mutual information carried by *that* collection about the target. But this is simply the mutual information of that collection:

$$I_\cap(T:\mathbf{a}_1) \overset{\text{Eq. 4.6}}{=} \sum_{f(\mathbf{a}_i)=1 \forall i=1,\ldots,m} \Pi(f) = \sum_{f(\mathbf{a}_1)=1} \Pi(f) \overset{\text{Eq. 4.5}}{=} I(T:\mathbf{a}_1) \tag{4.7}$$

Accordingly, mutual information has been called "self-redundancy" in the PID literature (although not based on parthood arguments) [19]. The relationship between redundant information and atoms is as follows: Only the "all-way" redundancy, i.e. the information shared by *all* n sources is itself an atom. Any other redundancy, such as the redundancy of only a subset of sources, is a composite quantity made up out of multiple atoms.

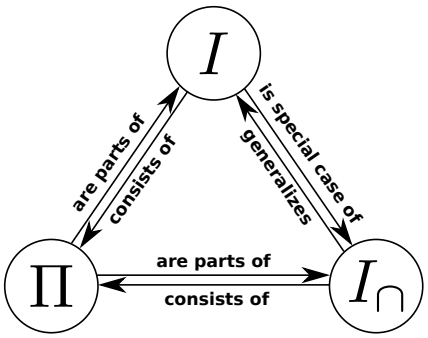

Fig. 4.4: Relationships between mutual information, redundant information, and information atoms.

Show that a measure of redundant information leads to a unique solution for the information atoms

There is a very useful fact about parthood distributions that will help us to obtain a unique solution for the atoms given an appropriate measure of redundant information: parthood distributions can be ordered in a very natural way into a lattice structure that is tightly linked to the idea of redundancy. The lattice for the case of three sources is shown in Figure 4.5. The parthood distributions are ordered as follows: If there is a 1 in certain positions on a parthood distribution f, then all the parthood distributions g below it also have a 1 in the same positions, plus some additional ones. Or in terms of the atoms corresponding to these parthood distributions: If an atom $\Pi(f)$ is part of the information provided by some collections of sources, then all the atoms $\Pi(g)$ below it are also part of the information provided

by these collections. Formally, we will denote this ordering by $\sqsubseteq$ and it is defined as

$$f \sqsubseteq g \Leftrightarrow (f(\mathbf{a}) = 1 \rightarrow g(\mathbf{a}) = 1 \text{ for any } \mathbf{a} \subseteq \{1, \ldots, n\}) \tag{4.8}$$

For n information sources we will denote the lattice of parthood distributions by $(\mathcal{B}_n, \sqsubseteq)$, where $\mathcal{B}_n$ is the set of all parthood distribution in the context of n sources (for proof that this structure is in fact a lattice in the formal sense see Appendix 4.8.2.

Note that the different "levels" of the lattice contain parthood distributions with the same number of ones and that higher level parthood distributions contain *less* ones: At the very top in Figure 4.5, there is the parthood distribution describing the atom that is *only* part of the joint mutual information provided by all three sources combined, i.e. the synergy of the three sources. One level down, there are the three parthood distributions that assign the value 1 exactly two times. Yet another level down, we find the three possible parthood distributions that assign the value 1 exactly three times. And so on and so forth until we reach the bottom of the lattice which corresponds to the information shared by all three sources. Accordingly the corresponding parthood distribution assigns the value 1 to all collections (except of course the empty collection).

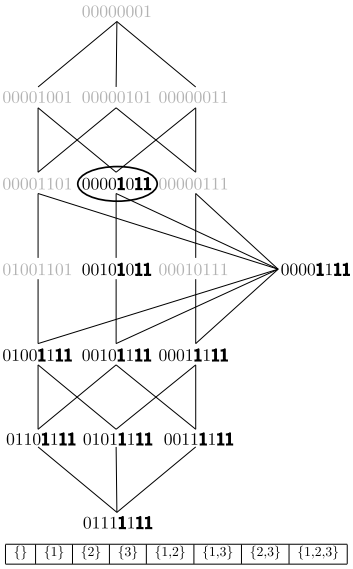

Fig. 4.5: Lattice of parthood distributions for the case of three information sources. The parthood distributions are represented as bit-strings where the i-th bit is the value that the parthood distribution assigns to the i-th collection of sources. The order of these collections is shown below the lattice for reference. A distribution f is below a distribution g just in case f has value 1 in the same positions as g and in some additional positions. This is illustrated for the parthood distribution highlighted by the black circle. The positions in which it assigns the value 1 are marked in bold face.

Ordering all the parthood distributions (and hence atoms) into such a lattice provides a good overview that tells us how many atoms exist for a given number of source variables and what their characteristic parthood relationships are. But the lattice plays a much more profound role because it is very closely connected to the concept of redundant information. The idea is to associate with each parthood distribution in the lattice a particular redundancy: the redundant information of all the collections that are assigned the value 1 by the distribution. In other words, for any parthood distribution f we consider the redundancy

$$I_\cap(T : f) := I_\cap\left(T : (\mathbf{a} \mid f(\mathbf{a}) = 1)\right) \tag{4.9}$$

For example, in the case of three sources, the redundant information associated with the parthood distribution that assigns value 1 to collections $\{1,2\}$, $\{2,3\}$, and $\{1,2,3\}$, and value 0 to all other collections (the one emphasized in Figure 4.5), is simply $I_\cap(T : \{1,2\},\{2,3\},\{1,2,3\})$. We saw in the previous section that any redundancy $I_\cap(T : \mathbf{a}_1,\ldots,\mathbf{a}_m)$ is the sum of all atoms that are part of the information provided by each of the $\mathbf{a}_i$. Now here is the connection between the lattice and redundant information: these atoms are the ones that have value 1 on each $\mathbf{a}_i$. But, by definition of the ordering, these are precisely the ones corresponding to parthood distributions *below and including* the parthood distribution for which we are computing the associated redundancy. In other words, the redundant information associated with a parthood distribution f can always be expressed as

$$I_\cap(T : f) = \sum_{g \sqsubseteq f} \Pi(g) \tag{4.10}$$

In this way we obtain one equation per parthood distribution. And since there are as many information atoms as parthood distributions, we obtain as many equations as unknowns. This is already a good sign. But is a unique solution for the information atoms guaranteed? This question can be answered affirmatively by noting that the system of equations described by (4.10) (one equation per f) is not just any linear system, but has a very special structure: one function $I_\cap(T : f)$ evaluated at a point f on a lattice is the sum of another function $\Pi(f)$ over all points on the lattice below and including the point f. The process of solving such a system for the $\Pi(f)$'s once all the $I_\cap(T : f)$'s are given, or in other words *inverting* equation (4.10), is called *Moebius Inversion*. Crucially, a unique solution is guaranteed for any real or even complex valued function $I_\cap$ that we may put on the lattice [113].

This means that we have now completely shifted the problem of determining the sizes of the information atoms to the problem of coming up with a reasonable definition of redundant information $I_\cap(T : f)$. Even though we have to define

this quantity for *each* parthood distribution f this is still a much simpler task. The reason is that all the $I_\cap$'s represent exactly the same *type* of information, namely redundant information. On the other hand, the information atoms Π represent completely different types of information. Even in the simplest case of two sources we have to deal not only with redundant information, but also unique information and synergistic information. And the story gets more and more complicated the more information sources are considered.

Now, note that apparently we only need to define quite special redundant information terms, namely the redundancies associated with parthood distributions $I_\cap(T : f)$ (see definition (4.9)). However, we will now show that these are in fact *all* possible redundancies, i.e. the redundancy of any tuple of collections of sources $\mathbf{a}_1, \ldots, \mathbf{a}_m$ is necessarily equal to a redundancy associated with a specific parthood distribution. The reason for this is that the quantitative relation between atoms and redundant information (equation (4.6)) not only provides a way to solve for the information atoms once we know $I_\cap$, it also implies that $I_\cap$ has to satisfy the following invariance properties:

1. $I_\cap(T : \mathbf{a}_1, \ldots, \mathbf{a}_m) = I_\cap(T : \mathbf{a}_{\sigma(1)}, \ldots, \mathbf{a}_{\sigma(m)})$ for any permutation σ (**symmetry**)

2. If $\mathbf{a}_i = \mathbf{a}_j$ for $i \neq j$, then $I_\cap(T : \mathbf{a}_1, \ldots, \mathbf{a}_m) = I_\cap(T : \mathbf{a}_1, \ldots, \mathbf{a}_{i-1}, \mathbf{a}_{i+1}, \ldots, \mathbf{a}_m)$ **(idempotency)**

3. If $\mathbf{a}_i \supset \mathbf{a}_j$ for $i \neq j$, then $I_\cap(T : \mathbf{a}_1, \ldots, \mathbf{a}_m) = I_\cap(T : \mathbf{a}_1, \ldots, \mathbf{a}_{i-1}, \mathbf{a}_{i+1}, \ldots, \mathbf{a}_m)$ **(invariance under superset removal / addition)**

4. $I_\cap(T : \mathbf{a}) = I(T : \mathbf{a})$ (**self-redundancy**)

We can easily ascertain that any measure of redundant information $I_\cap$ has to have these properties by taking a closer look at the condition describing which atoms to sum over in order to obtain a particular redundant information term $I(T : \mathbf{a}_1, \ldots, \mathbf{a}_m)$: we have to sum over the atoms with parthood distribution satisfying $f(\mathbf{a}_i) = 1$ for all $i = 1, \ldots, m$. Now whether or not this condition is true of a given parthood distribution f, first, does not depend on the *order* in which the collections $\mathbf{a}_i$ are given (symmetry), secondly, it does not depend on whether the same collection $\mathbf{a}$ is repeated multiple times (idempotency), and thirdly, it does not matter whether we add or remove some collection $\mathbf{a}_i$ that is a proper superset of some other collection (superset removal/addition). This fact is due to the monotonicity constraint on parthood distributions. Finally, the "self-redundancy" property was already established in the previous section.

These invariance properties are referred in the literature as the Williams and Beer axioms for redundant information [91] (in addition there is a *quantitative* monotonicity axiom that we reject. See §4.6). However, in the parthood formalism described here they are not themselves axioms but are *implied* by the core principles we have set out. The first two invariance properties imply that we may restrict ourselves to *sets* instead of tuples of collections in defining $I_\cap$. The third constraint additionally tells us that we can restrict ourselves to those sets of collections $\{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ such that no collection $\mathbf{a}_i$ is a superset of another collection $\mathbf{a}_j$. Such sets of collections are called *antichains*. Hence, the redundancy of *any* tuple of collections of sources $\mathbf{a}_1, \ldots, \mathbf{a}_m$ is necessarily equal to the redundancy associated with a particular antichain. This antichain results from ignoring the order and repetitions of the $\mathbf{a}_i$, and removing any supersets. For instance, $I_\cap(T : \{1\}, \{1\}, \{2\}, \{1, 2\}) = I_\cap(T : \{1\}, \{2\})$.

We can now see that the redundancies $I_\cap(T : f)$ are in fact all possible redundancies by associating with any antichain $\alpha = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ a parthood distribution f_α that assigns the value 1 to all $\mathbf{a}_i$ *and all supersets of these collections*, while it assigns the value 0 to all other collections. Now, due to the invariance of $I_\cap$ under removal of supersets, it immediately follows that $I_\cap(T : f_\alpha) = I_\cap(T : \alpha)$. So in conclusion, there is one redundancy for each antichain α and these redundancies are equal to the redundancies associated with the corresponding parthood distributions. Hence the redundancies $I_\cap(T : f)$ are in fact *all* possible redundancies.

Of course, there is also an inverse mapping associating with any parthood distribution f an antichain α_f. In fact, the lattice of parthood distributions $(\mathcal{B}_n, \sqsubseteq)$ is *isomorphic* to a lattice of antichains $(\mathcal{A}_n, \preceq)$ equipped with an ordering relationship that was originally introduced by Crampton and Loizou [114] and used by Williams and Beer in their original exposition of PID. The formal proof of this fact is postponed to section §4.4 where a third perspective on PID, the logical perspective, is introduced.

In the next section, we will tackle the problem of defining a measure of redundant information for each parthood distibution / antichain by connecting PID theory to formal logic. The measure $I_\cap^{\text{sx}}$ derived in this way is identical to the one described in [115]. In showing how this measure can be inferred from logical- and parthood-principles we aim to 1) strengthen the argument for $I_\cap^{\text{sx}}$, and 2), open the gateway between PID-theory and formal logical. This latter point is elaborated in §4.4.

4.3 Using logic to derive a measure of redundant information

We have now solved the PID problem up to specifying a reasonable measure of redundant information $I_\cap$ between collections that form an antichain. In this section, we will derive such a measure. In doing so we will first move from the level of random variables $T, S_1, \ldots, S_n$ to the level of particular realizations $t, s_1, \ldots, s_n$ of these variables. This level of description is generally called the *pointwise* level and has been used as the basis of classical information theory by Fano [35]. Pointwise approaches to PID have been put forth by [91] and [115].

Note that moving to the level of realizations simplifies the problem considerably because realizations are much simpler objects than random variables. A realization is simply a specific symbol or number whereas a random variables is an object that may take on various different values and can only be fully described by an entire probability distribution over these values.

4.3.1 Going Pointwise

The idea underlying the pointwise approach is to consider the information provided by a particular joint realization (observation) of the source random variables about a particular realization (observation) of the target random variable. So from now on we assume that these variables have taken on *specific* values $s_1, \ldots, s_n, t$. It was shown by Fano [35] that the whole of classical information theory can be derived from this pointwise level. By placing a certain number of reasonable constraints or axioms on pointwise information, it follows that this information must have a specific form. In particular, the pointwise mutual information $i(t : s)$ is given by

$$i(t : s) := \log \left(\frac{P(t|s)}{P(t)} \right) \tag{4.11}$$

The mutual information $I(T : S)$ is then simply defined as the *average* pointwise mutual information. Note that pointwise mututal information (in contrast to mutual information) can be both positive and negative. It essentially measures whether we are guided in the right or wrong direction with the respect to the actual target realization t. If the conditional probability of $T = t$ given the observation of $S = s$ is larger than the unconditional (prior) probability of $T = t$, then we are guided in the right direction: The actual target realization is in fact t and observing that $S = s$ makes us more likely to think so. Accordingly, in this case the pointwise mutual

information is *positive*. On the other hand, if the conditional probability of $T = t$ given the observation of $S = s$ is smaller than the unconditional (prior) probability of $T = t$, then we are guided in the wrong direction: Observing $S = s$ makes us less likely to guess the correct target value. In this case the pointwise mutual information is *negative*. The joint pointwise mutual information of source realizations $s_1, \ldots, s_n$ about the target realization is defined in just the same way:

$$i(t : s_1, \ldots, s_n) := \log \left(\frac{P(t|s_1, \ldots, s_n)}{P(t)} \right) \tag{4.12}$$

The idea is now to perform the entire partial information decomposition on the pointwise level, i.e. to decompose the pointwise joint mutual information $i(t : s_1, \ldots, s_n)$ that the source realizations provide about the target realization [91]. This leads to *pointwise atoms* $\pi_{s_1,\ldots,s_n,t}$ (in the following we will generally drop the subscript). Crucially, we are only changing the quantity to be decomposed from $I(T : S_1, \ldots, S_n)$ to $i(t : s_1, \ldots, s_n)$. Otherwise, the idea is completely analogous to what we have discussed in §4.2 (simply replace I by i and Π by π): the goal is to decompose the pointwise mutual information into information atoms that are characterized by their parthood relations to the pointwise mutual information provided by the different possible collections of source realizations. These atoms have to stand in the appropriate relationship to *pointwise redundancy*: the pointwise redundancy $i_\cap(t : \mathbf{a}_1, \ldots, \mathbf{a}_m)$ is the sum of all pointwise atoms $\pi(f)$ that are part of the information provided by *each* collection of source realizations $\mathbf{a}_i$. By exactly the same argument as described in §4.2ciii, there is a unique solution for the pointwise atoms once a measure of pointwise redundancy $i(t : \alpha)$ is fixed for all antichains $\alpha = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$. The *variable-level* atoms Π are then defined as the *average* of the corresponding pointwise atoms:

$$\Pi(f) = \sum_{s_1,\ldots,s_n,t} P(s_1, \ldots, s_n, t) \pi_{s_1,\ldots,s_n}(f) \tag{4.13}$$

We are now left with defining the pointwise redundancy $i_\cap$ among collections of source realizations. As noted above this is a much easier problem than coming up with a measure of redundancy among collections of entire source variables. In the next section, we show how the pointwise redundancy of multiple collections of source realizations can be expressed as the information provided by a particular *logical statement* about these realizations.

4.3.2 Defining pointwise redundancy in terms of logical statements

The language of formal logic allows us to form statements about the source realizations. In particular, we will consider statements made up out of the following ingredients:

1. n basic statements of the form $S_i = s_i$, i.e. "Source S_i has taken on value s_i"
2. the logical connectives $\wedge$ (and), $\vee$ (or), $\neg$ (not), $\rightarrow$ (if, then)
3. brackets),(

In this way, we may form statements such as $S_1 = s_1 \wedge S_2 = s_2$ ("Source S_1 has taken on value s_1 and source S_2 has taken on value s_2") or $S_1 = s_1 \vee (S_2 = s_2 \wedge S_3 = s_3)$ ("Either source S_1 has taken on value s_1 or source S_2 has taken on value s_2 and source S_3 has taken on value s_3"). Now we may ask: What is the information provided by the truth of such statements about the target realization t? Classical information theory allows us to quantify this information as a pointwise mutual information: Let A be any statement of the form just described, then the information $i(t : A)$ provided by the truth of this statement is

$$i(t : A) := i(t : \mathbb{I}_A = 1) = \log \left(\frac{P(t|A \text{ is true})}{P(t)} \right) \tag{4.14}$$

where $\mathbb{I}_A$ is the *indicator random variable* of the event that the statement A is true, i.e. $\mathbb{I}_A = 1$ if the event occurred and $\mathbb{I}_A = 0$ if it did not. The interpretation of this information is that it measures whether and to what degree we are guided in the right or wrong direction with respect to the actual target value once we learn that statement A is true.

Note that according to this definition the pointwise mutual information provided by a collection of source realizations $i(t : \mathbf{a})$ is the information provided by the truth of the *conjunction* $\bigwedge_{i \in \mathbf{a}} S_i = s_i$:

$$i(t : \mathbf{a}) = i \left(t : \bigwedge_{i \in \mathbf{a}} S_i = s_i \right) \tag{4.15}$$

Therefore, the information redundantly provided by collections of source realizations $\mathbf{a}_1, \ldots, \mathbf{a}_m$ is the information redundantly provided by the truth of the corresponding conjunctions. Now, what is this information? We propose that in general the following principle describes redundancy among statements:

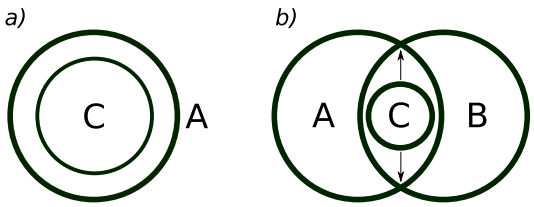

Fig. 4.6: (a) Information diagram depicting the information provided by statements A and C. If statement C is logically weaker than statement A, i.e. if C is implied by A, then the information provided by C has to be part of the information provided by A. (b) Information diagram depicting the information provided by statements A, B, and C. C is assumed to be logically weaker than both A and B. Thus it has to be part of the information provided by A and also part of the information provided by B. Accordingly, it is contained in the "overlap", i.e. the redundant information of A and B. In order to obtain the entire redundant information statement C has to be "maximized", i.e. it has to be chosen as the strongest statement weaker than both A and B (this is indicated by the arrows).

Core Principle 4. *The information redundantly provided by the truth of the statements $A_1, \ldots, A_m$ is the information provided by the truth of their disjunction $A_1 \vee \ldots \vee A_m$.*

There are two motivations for this principle: First, the logical inferences to be drawn from the disjunction $A \vee B$ are precisely the inferences that can be drawn *redundantly* from both A and B. If some conclusion C logically follows from both A and B, then it also follows from $A \vee B$. Conversely, any conclusion C that follows from the disjunction $A \vee B$ follows from both A and B. Formally,

$$A \vee B \models C \Leftrightarrow A \models C \text{ and } B \models C \tag{4.16}$$

where $\models$ denotes logical implication. The second motivation again invokes the idea of parthood relationships: *If some statement C is logically weaker than a statement A, then the information provided by C should be part of the information provided by A.* For instance, the information provided by the statement $S_1 = s_1$ has to be part of the information provided by the statement $S_1 = s_1 \wedge S_2 = s_2$. This idea is illustrated in the information diagram on the left side in Figure 4.6.

Now, this idea implies that if a statement C is weaker than *both* A and B, then the information provided by C is part of the information carried by A and also part of the information carried by B. But this means that the information provided by C is part of the *redundant information* of A and B. In order to obtain the *entire* redundant information, the statement C should therefore be chosen as the *strongest* statement logically weaker than both A and B (see right side of Figure 4.6). But this statement is the disjunction $A \vee B$ (or any equivalent statement).

Based on these ideas we can now finally formulate our proposal for a measure of pointwise redundancy $i_\cap(t : \mathbf{a}_1, \ldots, \mathbf{a}_m)$. We noted above that the information redundantly provided by collections of realizations $\mathbf{a}_1, \ldots, \mathbf{a}_m$ is the information redundantly provided by the conjunctions $\bigwedge_{i \in \mathbf{a}_j} S_i = s_i$. And by the arguments just presented this is the information provided by the *disjunction of these conjunctions*. We denote the function that measures pointwise redundant information in this way by $i_\cap^{\text{sx}}$ (for reasons that will be explained shortly). It is formally defined as:

$$i_\cap^{\text{sx}}(t : \mathbf{a}_1, \ldots, \mathbf{a}_m) := i\left(t : \bigvee_{j=1}^{m} \bigwedge_{i \in \mathbf{a}_j} S_i = s_i\right) \tag{4.17}$$

Recall that by definition this is the pointwise mutual information provided by the truth of the statement in question. Hence, it measures whether and to what degree we are guided in the right or wrong direction with respect to the actual target value t once we learn that the statement is true.

We have now arrived at a *complete* solution to the partial information decomposition problem: Given the measure $i_\cap^{\text{sx}}$ we may carry out the Moebius-Inversion

$$i_\cap^{\text{sx}}(t : f) = \sum_{g \sqsubseteq f} \pi^{\text{sx}}(f) \tag{4.18}$$

in order to obtain the pointwise atoms π^{sx}. This has to be done for *each* realization $s_1, \ldots, s_n, t$. The obtained values can then be averaged as per Equation (4.13) to obtain the variable-level atoms Π^{sx}.

As shown in [115], the measure $i_\cap^{\text{sx}}$ can also be motivated in terms of the notion of *shared exclusions* (hence the superscript "sx"). The underlying idea is that redundant information is linked to possibilities (i.e. points in sample space) that are redundantly excluded by multiple source realizations. We argue that the fact that the measure $i_\cap^{\text{sx}}$ can be derived in these two independent ways provides further support for its validity. We offer a freely accessible implementation of the $i_\cap^{sx}$ PID as part of the IDTxl toolbox [103]. Worked examples of its computation and details on the computational complexity can be found in [115].

In the following section, we show that the value of formal logic within the theory of partial information decomposition goes far beyond helping us to define a measure of pointwise redundant information. In fact, similar to the lattices of parthood distributions and antichains, there is a lattice of logical statements that can equally be used as the basic mathematical structure of partial information decomposition. This lattice is particularly useful because the ordering relationship turns out to be

very simple and well-understood: the relation of logical implication. We will show that this perspective also offers an independent starting point for the development of PID theory.

4.4 The logical perspective

4.4.1 Logic Lattices

The considerations of the previous section identified the information redundantly provided by collections $\mathbf{a}_1, \ldots, \mathbf{a}_m$ with the information provided by a particular logical statement: a disjunction of conjunctions of basic statements of the form $S_i = s_i$. This has an interesting implication: there is a one-to-one mapping between antichains α and logical statements. Let us now look at this situation a bit more abstractly by replacing the concrete statements $S_i = s_i$ with *propositional variables* $\varphi_1, \ldots, \varphi_n$. Together with the logical connectives $\neg, \vee, \wedge, \rightarrow$ (plus brackets) these form a language of propositional logic [116]. We will denote this language by $\mathbb{L}$. We may now formally introduce a mapping Ψ from the set of antichains $\mathcal{A}$ into $\mathbb{L}$ via

$$\Psi : \mathcal{A} \to \mathbb{L}, \quad \text{where } \alpha \rightarrowtail \tilde{\alpha} := \bigvee_{a \in \alpha} \bigwedge_{i \in a} \varphi_i \tag{4.19}$$

In other words, α is mapped to a statement by first conjoining the φ_i corresponding to indices *within* each $\mathbf{a}_i$ and then disjoining these conjunctions. For instance, the antichain $\{\{1,2\},\{2,3\}\}$ will be associated with the statement $(\varphi_1 \wedge \varphi_2) \vee (\varphi_2 \wedge \varphi_3)$. Note that if we interpret the propositional variables φ_i as "source S_i has taken on value s_i", then this is of course precisely the mapping of an antichain to the statement providing the redundant information (in the sense of $i_\cap^{sx}$) associated with that antichain. [1]

The range $\mathcal{L} \subseteq \mathbb{L}$ of Ψ is *set of all disjunctions of logically independent conjunctions of pairwise distinct propositional variables*. The logical independence of the conjunctions is the logical counterpart of the antichain property. The "pairwise distinct" condition ensures that the same atomic statement does not occur multiple times in any conjunction. The set $\mathcal{L}$ can now be equipped with the relationship of logical implication $\models$ in order to obtain a new structure $(\mathcal{L}, \models)$ which we will show to be

[1] There is a slight ambiguity in the definition of Ψ since the *order* of the conjunctions $\bigwedge_{i \in a} \varphi_i$ and statements φ_i is not specified. This problem can be solved, however, by choosing any enumeration of the elements $\mathbf{a}$ of the powerset of $\{1, \ldots, n\}$ and ordering the conjunctions $\bigwedge_{i \in \mathbf{a}} \varphi_i$ accordingly. The propositional variables φ_i within the conjunctions may simply be ordered by ascending order of their indices.

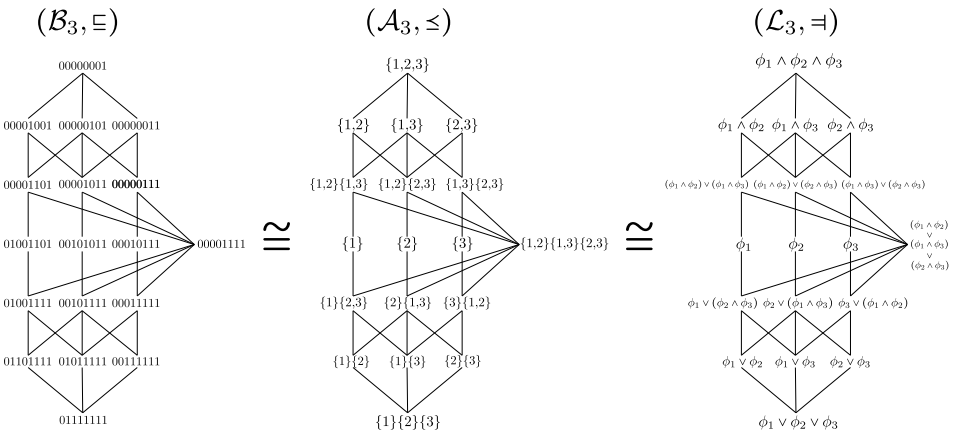

Fig. 4.7: The three isomorphic worlds of partial information decomposition: parthood distributions, antichains, and logical statements.

isomorphic to the lattices of antichains and parthood distributions. Here $\models$ means "implies" and $\dashv$ means "is implied by".

Based on these concepts, the following theorem expresses the isomorphism of $(\mathcal{L}, \dashv)$ to the lattices of antichains and parthood distributions:

Theorem 9. *For all $n \in \mathbb{N}$: $(\mathcal{L}_n, \dashv)$ is isomorphic to $(\mathcal{A}_n, \preceq)$ and $(\mathcal{B}_n, \sqsubseteq)$*

Proof. See Electronic Appendix 4.8.2. □

Corollary 1. *For all $n \in \mathbb{N}$: $(\mathcal{L}_n, \dashv)$ is a poset and specifically a lattice.*

In this way the logical perspective is put on equal footing with the parthood perspective and "antichain" perspective described by Williams and Beer [19]. They are in fact three equivalent ways to describe the mathematical structure underlying partial information decomposition. These three "worlds" of PID are illustrated in Figure 4.7 for the case of three information sources.

Intuitively, the logic lattice can be understood as a hierarchy of logical constraints describing how (i.e. via which collections of sources) information about the target may be accessed. The information atom associated with a node $\tilde{\alpha}$ in the logic lattice is *exactly* the information about the target that can be accessed in the way described by the constraint $\tilde{\alpha}$. For example, the information shared by all sources $\Pi(\{1\}, \{2\}, \{3\})$ is to be found at the very bottom of the logic lattice because access to this information is constrained in the least possible way: the shared information can be accessed via *any* source (i.e. via source 1 *or* source 2 *or* source 3). By monotonicity, the shared information is of course also accessible via any *collection*

of sources so that in total there are seven ways to access it (one per collection). By contrast, the all-way synergy $\Pi(\{1,2,3\})$ is located at the very top of the logic lattice because access to it is most heavily constrained: the synergy can only be accessed if all sources are known at the same time. Hence, there is only a single way (collection) to access it. All other atoms are to be found in between these two extremes. For instance, the atom corresponding to the constraint $\varphi_1 \vee (\varphi_2 \wedge \varphi_3)$ is exactly the information that can be accessed either via source 1 or via sources 2 and 3 jointly (and of course via any superset of these collections by monotonicity) *but not in any other way* (i.e. not via sources 2 or 3 individually). So in total there are five ways to access it corresponding to the collections $\{1\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}$. In general, the atoms on the k-th level of the logic lattice (starting to count at the top) are precisely the atoms that can be accessed via k collections of sources (compare this to the very similar insight in §4.2ciii).

Finally, one may also associate a redundant information term with each node in the logic lattice by interpreting the statements as *sufficient conditions* for access instead of *constraints*, i.e. sufficient and necessary conditions, on access. For instance, the redundancy associated with the statement $\varphi_1 \wedge \varphi_2 \wedge \varphi_3$ would be all information for which joint knowledge of all three sources is sufficient. But this is of course *all* information contained in the sources, i.e. the entire joint mutual information. By contrast, the information atom associated with the same statement is the information for which joint knowledge of all three sources is not only sufficient but also necessary, i.e. it cannot be obtained via any other collection of sources. Or put generally: while the redundancy is the information we obtain *if* we have knowledge of certain collections of sources, the information atom is the information we obtain *if and only if* we have such knowledge. Defined in this way the redundant information of a lattice node is again the sum of atoms associated with nodes below and including it.

In this way the logical perspective can be used as an independent starting point to develop PID theory. Instead of characterizing atoms by their defining parthood relations one might equally characterize them by their defining access constraints and relate them to the notion of redundant information in the way just described. This is summarized in the following Core Principle:

Core Principle 5. *Each atom of information is characterized by a logical constraint describing via which collections of sources it can be accessed. The atom* $\Pi(\tilde{\alpha})$ *associated with constraint* $\tilde{\alpha} = \bigvee_{\mathbf{a}\in\alpha} \bigwedge_{i\in\mathbf{a}} \varphi_i$ *is exactly that part of the joint mutual information about the target that can be accessed if and only if we have knowledge of any one of the collections of sources* $\mathbf{a}$.

Now that we have fully introduced both the parthood and logical approaches to PID it is worth noting their key difference to the original "antichain" approach by Williams and Beer: whereas the parthood and logic approaches are looking at the problem from the perspective of the atoms and seek to describe their defining parthood relations / access constraints, the antichain based approach starts off by placing certain axioms on measures of redundant information leading to the insight that the definition of redundancy may be restricted to antichains. The atoms are then *indirectly* introduced in terms of a Moebius-Inversion over the lattice of antichains.

The next section highlights an additional use of logic lattices, namely as a mathematical tool to analyse the structure of PID lattices.

4.4.2 Using logic lattices as a mathematical tool to analyse the structure of PID lattices

One advantage that logic lattices have over the lattices of antichains and parthood distributions is that their ordering relationship is particularly natural and well-understood: logical implication between statements. By contrast, the ordering relation $\preceq$ on the lattice of antichains only seems to have been studied in quite restricted order theoretic contexts so far. Furthermore, it is a purely technical concept that does not have a clear-cut counterpart in ordinary language. Because of the simplicity of its ordering relation, many important order theoretic concepts have a simple interpretation within the logic lattice. This makes it a useful tool to understand the structure of the lattice itself which in turn is relevant to the computation of information atoms.

There is an interesting fact about the statements in $\mathcal{L}$ that will be useful in the following investigations: they correspond to statements with monotonic truth-tables. The truth-table $T_{\tilde{\alpha}} : \mathcal{V} \to \{0, 1\}$ of a statement $\tilde{\alpha}$ describes which models $V \in \mathcal{V}$ satisfy $\tilde{\alpha}$ ("make $\tilde{\alpha}$ true"), i.e.

$$T_{\tilde{\alpha}}(V) = \begin{cases} 1 & \text{if } \models_V \tilde{\alpha} \\ 0 & \text{otherwise} \end{cases} \tag{4.20}$$

A truth-table T shall be called *monotonic* just in case $\forall i \in \{1, \ldots, n\}$

$$(V(\varphi_i) = 1 \to V'(\varphi_i) = 1) \Rightarrow \big(T(V) = 1 \to T(V') = 1\big) \tag{4.21}$$

In other words, suppose a statement $\tilde{\alpha}$ is satisfied by a valuation V. Now suppose further that a new valuation V' is constructed by flipping one or more zeros to one

in V. Then $\tilde{\alpha}$ has to be satisfied by V' as well. Making some φ_i true that were previously false cannot make $\tilde{\alpha}$ false if it was previously true. With this terminology at hand the following proposition can be formulated:

Proposition 8. *All $\tilde{\alpha} \in \mathcal{L}$ have monotonic truth-tables. Conversely, for all monotonic truth-tables T, there is exactly one $\tilde{\alpha} \in \mathcal{L}$ such that $T_{\tilde{\alpha}} = T$. In other words, the statements in $\mathcal{L}$ are, up to logical equivalence, exactly the statements of propositional logic with monotonic truth-tables.*

Proof. See Appendix 4.8.3 □

Now, it was shown in [91] that the information atoms have a closed form solution in terms of the *meets* of any subset of children of the corresponding node in the lattice. The *meet* (infimum) and *join* (supremum) operations, however, have quite straightforward interpretations on $(\mathcal{L}, \dashv)$: The meet of two statements $\tilde{\alpha}$ and $\tilde{\beta}$ is the strongest statement logically weaker than both $\tilde{\alpha}$ and $\tilde{\beta}$. Similarly, the join is the weakest statement logically stronger than both $\tilde{\alpha}$ and $\tilde{\beta}$. The meet is logically equivalent (though not identical) to the *disjunction* of $\tilde{\alpha}$ and $\tilde{\beta}$ while the join is logically equivalent (though not identical) to their *conjunction*. The conjunction and disjunction of two elements of $\mathcal{L}$ do generally not lie in $\mathcal{L}$ because they do not necessarily have the appropriate form (disjunction of logically independent conjunctions). However, this can easily be remedied because both the disjunction and the conjunction of elements of $\mathcal{L}$ have monotonic truth-tables. Thus, by Proposition 8 there is a unique element in $\mathcal{L}$ with the same truth-table in both cases. These elements are therefore the meet and join. The detailed construction of meet and join operators is presented in Appendix 4.8.3.

Let us now turn to the notions of child and parent. A *child* of a statement $\tilde{\alpha} \in \mathcal{L}$ is a strongest statement strictly weaker than $\tilde{\alpha}$. Similarly, a *parent* of $\tilde{\alpha}$ is a weakest statement strictly stronger than $\tilde{\alpha}$. The following three propositions provide, first, a characterization of children in terms of their truth tables, second, a lower bound on the number of children of a statement, and third, an algorithm to determine all children of a statement. Due to the isomophism of antichains, parthood distributions, and logical statements, the propositions can be utilized to study any of these three structures.

Proposition 9 (Characterization of Children)**.** *$\tilde{\gamma} \in \mathcal{L}$ is a direct child of $\tilde{\alpha} \in \mathcal{L}$ if and only if $\tilde{\gamma}$ is true in all cases in which $\tilde{\alpha}$ is true plus exactly one additional case, i.e. just in case $T_{\tilde{\alpha}}(V) = 1 \rightarrow T_{\tilde{\gamma}}(V) = 1$ and $!\exists V \in \mathcal{V} : T_{\tilde{\gamma}}(V) = 1 \wedge T_{\tilde{\alpha}}(V) = 0$.*

Proof. See Appendix 4.8.3 □

Proposition 10 (Lower bound on number of children)**.** *Any $\alpha \in \mathcal{A}$ such that there is at least one $\mathbf{a} \in \alpha$ with $|\mathbf{a}| = k \geqslant 1$ has at least k children.*

Proof. See Appendix 4.8.3 □

Proposition 11 (Algorithm to determine children)**.** *The children of a statement $\tilde{\alpha}$ can be determined via the following algorithm (for a pseudocode version see Appendix 4.8.3). It proceeds in three steps:*

1. *Set k to the maximal number of ones occurring in a valuation that does not satisfy $\tilde{\alpha}$.*

2. *For each valuation V that does not satisfy $\tilde{\alpha}$ and contains k ones do the following:*

    a) *Check if there is a valuation with k+1 ones that does not satisfy $\tilde{\alpha}$ and results from flipping one or multiple zeros in V to one, i.e. a model V' such that $V(\varphi_i) = 1 \rightarrow V'(\varphi_i) = 1$. If there is such a valuation, then skip step b). Otherwise, proceed.*

    b) *Create a new monotonic truth-table by setting V to one, otherwise leaving the truth-table of $\tilde{\alpha}$ unchanged. The statement corresponding to this truth-table is a child of $\tilde{\alpha}$.*

3. *If $k > 0$, decrease k by 1 and repeat Step 2. Otherwise, terminate.*

Proof. See Appendix 4.8.3 □

This concludes our discussion of the relationship between formal logic and PID. In the next section we return to the parthood side of our story. In particular, we will address an apparent arbitrariness in the argument presented in §4.2c. Here we showed that the sizes of the atoms of information can be obtained once a measure of redundant information is specified. Now, one may ask of course: why redundant information? Couldn't the same purpose be achieved by utilizing some other informational quantity such as synergistic or unique information? We will now discuss how the parthood approach can help answering this question in a systematic way.

4.5 Non-Redundancy based PIDs

Let us briefly revisit the structure of the argument in §4.2c. It involved three steps (presented in slightly different order above): First, based on the very concept of redundant information, we phrased a condition describing which atoms are part of which redundancies (Core Principle 3). Secondly, we showed that this parthood criterion entails a number of contraints on the measure $I_\cap$. Finally, we showed that, as long as these constraints are satisfied, we obtain a unique solution for the atoms of information. There is actually a fourth step: We would have to check that the information decomposition satisfies the consistency equations relating atoms to mutual information terms (Equation 4.5). However, in the case of redundant information this condition is trivially satisfied due to the self-redundancy property. In other words, the consistency equations are themselves part of the system of equations used to solve for the information atoms.

In order to obtain an information decomposition based on a quantity other than redundant information, lets call it $I^*(T : \mathbf{a}_1, \ldots, \mathbf{a}_m)$, we may use precisely the same scheme:

1. Define a condition $\mathcal{C}(f : \mathbf{a}_1, \ldots, \mathbf{a}_m)$ on parthood distributions f describing which atoms $\Pi(f)$ are part of $I^*(T : \mathbf{a}_1, \ldots, \mathbf{a}_m)$ for any given tuple of collections of sources $\mathbf{a}_1, \ldots, \mathbf{a}_m$. This leads to a system of equations:

$$I^*(T : \mathbf{a}_1, \ldots, \mathbf{a}_m) = \sum_{\mathcal{C}(f:\mathbf{a}_1,\ldots,\mathbf{a}_m)} \Pi(f) \tag{4.22}$$

2. Analyse which constraints on $I^*(T : \mathbf{a}_1, \ldots, \mathbf{a}_m)$ (e.g. symmetry, idempotency, ...) are implied by this relationship.

3. Show that given a choice of $I^*(T : \mathbf{a}_1, \ldots, \mathbf{a}_m)$ that satisfies the constraints, a unique solution for all information atoms $\Pi(f)$ can be obtained.

4. Show that the solution satisfies the consistency equation (4.5) relating information atoms and mutual information terms.

Let us work through these steps in specific cases.

4.5.1 Restricted Information PID

Recall that the redundant information of multiple collections of sources is the information we obtain *if* we have access to any of the collections. Similarly, we can define the information "restricted by" collections of sources $\mathbf{a}_1, \ldots, \mathbf{a}_m$ as any information we obtain *only if* we have access to at least one of the collections. For instance, assuming $n = 2$, the information restricted by the first source consists of its unique information and its synergy with the second source. Both of these quantities can only be obtained if we have access to the first source.

Thus, in general the restricted information $I_{\text{res}}(T : \mathbf{a}_1, \ldots, \mathbf{a}_m)$ should consist of all the atoms that are *only* part of the information carried by some of the $\mathbf{a}_i$ *but not part of the information provided by any other collection of sources*. Thus the parthood condition $\mathcal{C}_{\text{res}}$ is given by

$$\mathcal{C}_{\text{res}}(f : \mathbf{a}_1, \ldots, \mathbf{a}_m) \Leftrightarrow (f(\mathbf{b}) = 1 \rightarrow \exists i : \mathbf{b} \supseteq \mathbf{a}_i) \tag{4.23}$$

and we obtain the relation

$$I_{\text{res}}(T : \mathbf{a}_1, \ldots, \mathbf{a}_m) = \sum_{\mathcal{C}_{\text{res}}(f : \mathbf{a}_1, \ldots, \mathbf{a}_m)} \Pi(f) \tag{4.24}$$

Just as in the case of redundant information, this relationship implies a number of invariance properties for I_{res}: it has to be symmetric, idempotent, and invariant under superset removal/addition allowing us again to restrict ourselves to the set of antichains. The analogue of the "self-redundancy" property is that the restricted information of a collection of singletons $I_{\text{res}}(T : \{i_1\}, \ldots, \{i_m\})$ is equal to the *conditional mutual information provided by their union* $\alpha_\cup = \bigcup_{j=1}^{m} \{i_j\}$ *conditioned on all other sources*. So if $\alpha = \{\{i_1\}, \ldots, \{i_m\}\}$ is a collection of singletons, then:

$$I_{\text{res}}(T : \alpha) = I\left(T : (S_i)_{i \in \alpha_\cup} | (S_j)_{j \in \alpha_\cup^C}\right) \tag{4.25}$$

This can be established using the chain rule for mutual information as detailed in Appendix 4.8.4. The next step is to show that we may obtain a unique solution for the information atoms once a measure of restricted information satisfying these conditions is given. This can be achieved in much the same way as for redundant information. The restricted information associated with an antichain α can be expressed as a sum of information atoms $\Pi(\beta)$ below and including α in a specific lattice of antichains $(\mathcal{A}, \preceq')$. This lattice is simply the dual (inverted version) of the antichain lattice $(\mathcal{A}, \preceq)$, i.e.

$$\alpha \preceq' \beta \Leftrightarrow \beta \preceq \alpha \tag{4.26}$$

Accordingly, a unique solution is guaranteed via Moebius-Inversion of the relationship

$$I_{\text{res}}(T : \alpha) = \sum_{\beta \preceq' \alpha} \Pi_{\text{res}}(\alpha) \tag{4.27}$$

As a final step we need to show that the resulting atoms stand in the appropriate relationships to mutual information terms. These relationships are given by the consistency equation (4.5). Again using the chain rule it can be shown that this equation is equivalent to a condition relating conditional mutual information to the information atoms:

$$I(T : \mathbf{a}) = \sum_{f(\mathbf{a})=1} \Pi(f) \quad \Leftrightarrow \quad I(T : \mathbf{a}|\mathbf{a}^C) = \sum_{f(\mathbf{a}^C)=0} \Pi(f) \tag{4.28}$$

Now consider any collection of source indices $\mathbf{a} = \{j_1, \ldots, j_m\}$, then we obtain

$$\begin{aligned} I\left(T : \mathbf{a}|\mathbf{a}^C\right) &\overset{Eq.(4.25)}{=} I_{\text{res}}(T : \{j_1\}, \ldots, \{j_m\}) && (4.29)\\ &\overset{Eq.(4.24)}{=} \sum_{f(\mathbf{b})=1 \to \exists i : \mathbf{b} \supseteq \{j_i\}} \Pi_{\text{res}}(f) && (4.30)\\ &= \sum_{f(\mathbf{a}^C)=0} \Pi_{\text{res}}(f) && (4.31) \end{aligned}$$

where the last equality follows because in the case of singletons the parthood condition $\mathcal{C}_{\text{res}}$ reduces to $f(\alpha_\cup^C) = 0$. This establishes that the resulting atoms satisfy the consistency condition and we obtain a valid PID. In the following section we will use the same approach to analyse the question of whether a synergy based PID is possible.

4.5.2 Synergy based PID

Note that the restricted information of multiple collections of sources stands in a direct correspondence to a weak form of synergy which we will denote by $I_{\text{ws}}(T : \mathbf{a}_1, \ldots, \mathbf{a}_m)$. This quantity is to be understood as *the information about the target we cannot obtain from any individual collection* $\mathbf{a}_i$. Accordingly, the parthood criterion is

$$\mathcal{C}_{\text{ws}}(f : \mathbf{a}_1, \ldots, \mathbf{a}_m) \Leftrightarrow (\forall i \in \{1, \ldots, m\} : f(\mathbf{a}_i) = 0) \tag{4.32}$$

But this information is of course the same as the information that we can get only if some *other* collections are known (except subcollections of course), i.e.

$$I_{\text{ws}}(T : \mathbf{a}_1, \ldots, \mathbf{a}_m) = I_{\text{res}}(T : (\mathbf{b} \mid \forall i \mathbf{b} \not\subseteq \mathbf{a}_i)) \tag{4.33}$$

Consider the case of two sources: the information we cannot get from source 2 alone, $I_{\text{ws}}(T : \{2\})$, is the same as the information we can get only if the first source is known, $I_{\text{res}}(T : \{1\})$: unique information of source 1 plus synergistic information.

Due to this correspondence, the argument presented above can also be used to show that a consistent PID can be obtained by fixing a measure I_{ws} of weak synergy. Once such a measure is given we can first translate it to the corresponding restricted information terms and then perform the Moebius inversion of Equation (4.27) (alternatively, the above argument could be redeveloped directly for I_{ws} with minor modifications)

Interestingly, if we associate with every antichain α in the lattice $(\mathcal{A}, \preceq)$ the corresponding $I_{\text{ws}}(T : \beta)$ (so that $I_{\text{res}}(T : \alpha) = I_{\text{ws}}(T : \beta)$), then the β form an isomorphic lattice but with a different ordering (see Figure 4.8). Just as the original antichain lattice this structure on the antichains has been introduced by Crampton and Loizou [114].

In the PID field a restricted version of this lattice (i.e. restricted to a certain subset of antichains) has been described by [117] and [118] under the name “constraint lattice”. This terms is also appropriate in the present context: Intuitively, if we move up the constraint lattice we encounter information that satisfies more and more constraints. First, all of the information in the sources ($I_{\text{ws}}(T : \varnothing)$). This is the case of no constraints. Then all the information that is not contained in a particular individual source ($I_{\text{ws}}(T : \{1\})$ and $I_{\text{ws}}(T : \{2\})$). And finally the information that is not contained in any individual source ($I_{\text{ws}}(T : \{1\}, \{2\})$) .

Most recently, the full version of the lattice (i.e. defined on all antichains) has been utilized by [119] to formulate a synergy centered information decomposition. They call the lattice *extended constraint lattice* and define "synergy atoms" S_∂ in terms of a Moebius-Inversion over it. The concept of synergy S^α utilized in this approach closely resembles what we have called weak synergy. However, the decomposition is *structurally different* from the type of decomposition discussed here and generally assumed in previous work on PID. Even though it leads to the same number of atoms, these atoms do not stand in the expected relationships to mutual information. For instance, in the 2-sources case, there is no pair of atoms that necessarily adds up to the mutual information provided by the first source and no such pair of atoms for the second source. The consistency equation (4.5) is not satisfied (except for the full set of sources). This means that synergy atoms S_∂ are not directly comparable to standard PID atoms Π. They represent different types of information.

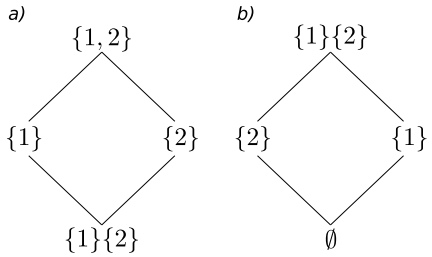

Fig. 4.8: (a) antichain lattice $(\mathcal{A}_2, \preceq)$ for two sources. Summing up the atoms *above* and including a node yields the restricted information of that node. (b) extended constraint lattice for two sources. The weak synergy associated with a node in the extended constraint lattice is the sum of atoms above and including the corresponding node in the left lattice. Note that following a widespread convention we left out the outer curly brackets around the antichains.

Let us now move towards stronger concepts of synergistic information. The reason for the term "weak" synergy is that a key ingredient of synergy seems to be missing in its definition: intuitively, the synergy of multiple sources is the information that cannot be obtained from any individual source but that becomes "visible" once we know all the sources at the same time. However, the definition of weak synergy only comprises the first part of this idea. The weak synergy $I_{\text{ws}}(T : \mathbf{a}_1, \ldots, \mathbf{a}_m)$ also contains parts that do not become visible even if we have access to *all* $\mathbf{a}_i$. For instance, given $n = 3$, the weak synergy $I_{\text{ws}}(T : \{1\}, \{2\})$ also contains the unique information of the third source $\Pi(\{3\})$ because this quantity is accessible from neither the first nor the second source.

So let us add this missing ingredient by strengthening the parthood criterion:

$$\mathcal{C}_{\text{ms}}(f : \mathbf{a}_1, \ldots, \mathbf{a}_m) \Leftrightarrow (\forall i \in \{1, \ldots, m\} : f(\mathbf{a}_i) = 0 \,\&\, f(\alpha_\cup) = 1) \tag{4.34}$$

We obtain a moderate type of synergy we denote by $I_{\text{ms}}(T : \mathbf{a}_1, \ldots \mathbf{a}_m)$. It has a nice geometrical interpretation: in an information diagram it corresponds to all atoms *outside* of all areas associated with the mutual information carried by some $\mathbf{a}_i$ but *inside* the area associated with the mutual information carried by the union of the $\mathbf{a}_i$ (see Figure 4.9). Furthermore, we can immediately see that the parthood condition cannot be satisfied for individual collections $\mathbf{a}$ (it demands $f(\mathbf{a}) = 0$ and $f(\mathbf{a}) = 1$ at the same time). This makes intuitive sense because the synergy of an individual collection appears to be an ill-defined concept: at least two things have to come together for there to be synergy. We will get back to the case of individual collections below.

Let us first see what properties are implied by $\mathcal{C}_{\text{ms}}$. It can readily be shown that I_{ms} is symmetric, idempontent, and invariant under *subset* removal. This again allows

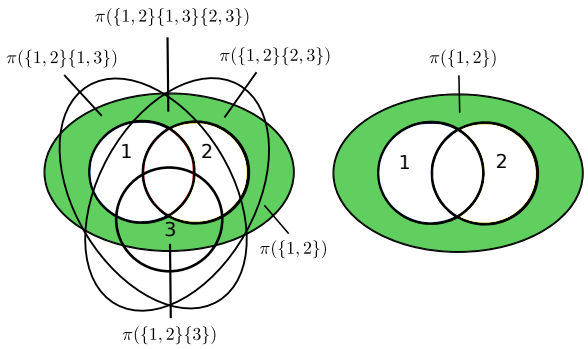

Fig. 4.9: Geometrical interpretation of moderate synergy $I_{\text{ms}}(T : \{1\},\{2\})$ for 2 and 3 sources.

us to restrict the domain of I_{ms} to the antichains. Additionally, I_{ms} satisfies the following condition:

$$\text{If } \exists i : \alpha_{\cup} = \mathbf{a}_i, \text{ then } I_{\text{ms}}(T : \alpha) = 0 \textbf{ (zero condition)} \tag{4.35}$$

This property says that whenever the union of the collection happens to be equal to one of collections then the moderate synergy must be zero. This is in particular the case for the moderate "self-synergy" of a single collection. On first sight this raises a problem since the synergy equations associated with individual collections become trivial ($0 = 0$) and do not impose any constraints on the atoms. This situation can be remedied, however, by noting that these missing constraints are provided by the consistency equations relating the atoms to mutual information / conditional mutual information. In this way a unique solution for the atoms is indeed guaranteed (one could also axiomatically set the “self-synergies” to the respective conditional mutual information terms). The proof of this statement is given in Appendix 4.8.4.

An instructive fact about the moderate synergy based PID is that the underlying system of equations does not have the structure of a Moebius-Inversion over a lattice: there is no arrangement of atoms into a lattice such that each $I_{\text{ms}}(T : \alpha)$ turns out to be the sum of atoms below and including a particular lattice node. The reason is that any finite lattice always has a unique least element. In other words, some atom would have to appear at the very bottom of the lattice and would therefore be contained in all synergy terms. However, in the case of moderate synergy, there is no such atom for $n \geqslant 3$. The only viable candidate would be the overall synergy $\Pi(\{1,\ldots,n\})$. But due to the condition that the synergistic information has to become visible if we know all collections in question, this atom is not contained e.g. in $I_{\text{ms}}(T : \{1\},\{2\})$.

Now one may wonder if the concept of synergy can be strengthened even further by demanding that the synergistic information should not be accessible from the *union of any proper subset* of the collections in question. For instance, the synergistic information $I_{\text{syn}}(T : \{1\}\{2\}\{3\})$ of sources 1, 2, and 3 should not be accessible from the collections $\{1, 2\}$, $\{1, 3\}$, or $\{2, 3\}$. We have to know *all three sources* to get access to their synergy. Thus, we may add this third constraint to obtain a strong notion of synergy we denote by $I_{\text{syn}}(T : \mathbf{a}_1, \ldots, \mathbf{a}_m)$. An atom $\Pi(f)$ should satisfy the corresponding parthood condition $\mathcal{C}_{\text{syn}}(f : \mathbf{a}_1, \ldots, \mathbf{a}_m)$ just in case

1. $f\left(\bigcup_{i=1}^{m} \mathbf{a}_i\right) = 1$
2. $\forall i \in \{1, \ldots, m\} : f(\mathbf{a}_i) = 0$
3. $\forall J \subset \{1, \ldots, m\}, |J| \geqslant 2 : \bigcup_{j \in J} \mathbf{a}_j \neq \bigcup_{i=1}^{m} \mathbf{a}_i \rightarrow f\left(\bigcup_{j \in J} \mathbf{a}_j\right) = 0$

The last condition is phrased as a conditional because the union of a proper subset of collection might happen to be equal to the union of all collections in question. Consider the case of three sources and the synergy $I_{\text{syn}}(T : \{1, 2\}\{1, 3\}\{2, 3\})$. In this case the union of a proper subset of these collections, for instance $\{1, 2\} \cup \{1, 3\}$, happens to be equal to the union of all $\mathbf{a}_i$.

Unfortunately, we do not obtain enough linearly independent equations to uniquely determine the atoms of information. This can be shown using the example of three sources. According to the parthood criterion, $I_{\text{syn}}(T : \{1\}\{2\}\{3\}) = \Pi(\{1, 2, 3\})$. But also $I_{\text{syn}}(T : \{1, 2\}\{1, 3\}\{2, 3\}) = \Pi(\{1, 2, 3\})$. This means that we do not obtain independent equations for each antichain. Or in linear algebras terms: our coefficient matrix will have two linearly dependent (actually identical) rows. Thus, a measure of strong synergy as described by $\mathcal{C}_{\text{syn}}$ cannot induce a unique PID.

4.5.3 Unique information PID

Let us briefly discuss the last obvious candidate quantity for determining the PID atoms: unique information [83]. The appropriate parthood criterion for a measure of unique information I_{unq} seems straightforward in the case of individual collections $\mathbf{a}$: It should consist of all atoms that are part of the information provided by the collection $\mathbf{a}$ but not part of the information provided by any other collection. This is what makes this information "unique" to the collection. Since there is always just one such atom this means that $I_{\text{unq}}(T : \mathbf{a}) = \Pi(\mathbf{a})$. For instance, $I_{\text{unq}}(T : \{1\}) = \Pi(\{1\})$, as expected. However, defining I_{unq} only for individual collections does not yield enough equations to solve for the atoms. We need one equation per antichain / parthood distribution, and hence, some notion of the unique information associated with *multiple* collections $\mathbf{a}_1, \ldots, \mathbf{a}_m$. This is a trickier question. What does it mean

for information to be unique to these collections? Certainly, uniqueness demands that this information should not be contained in any *other* collection. But what about the collections $\mathbf{a}_1, \ldots, \mathbf{a}_m$ themselves? It seems that the appropriate condition is that the unique information should consist of atoms that are contained in *all* of these collections. This idea aligns well with ordinary language: for instance, saying that a certain protein is unique to sheep and goats means that this protein is found in *both sheep and goats and nowhere else*. Using this idea, the parthood criterion becomes

$$\mathcal{C}_{\text{unq}}(f : \mathbf{a}_1, \ldots, \mathbf{a}_m) \Leftrightarrow (f(\mathbf{a}) = 1 \leftrightarrow \exists i : \mathbf{a} \supseteq \mathbf{a}_i) \tag{4.36}$$

However, this condition simply defines the atom $\Pi(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ making the unique information based PID possible but maybe not very helpful: it just amounts to defining all the atoms separately because $I_{\text{unq}}(T : \alpha) = \Pi(\alpha)$ for all antichains α.

4.6 Parthood descriptions vs. quantitative descriptions

Before concluding we would like to briefly point out an issue that arises quite naturally when thinking about information theory from a parthood perspective and that merits a few remarks: throughout this paper we have drawn a distinction between *parthood* relationships and *quantitative* relationships between information contributions. In particular, Core Principles 1 and 3 express parthood relationships between information atoms on the one hand and mutual information / redundant information on the other. Core Principle 2 by contrast describes the quantitative relationship between any information contribution and the parts it consists of. It is crucial to draw this distinction because these principles are logically independent. Consider the case of two sources: In this case, one could agree that the joint mutual information should consist of four parts while disagreeing that it should be the *sum* of these parts. On other hand, one could agree that the joint mutual information should be the sum of its parts but disagree that it consists of four parts.

The distinction between parthood relations and quantitative relations is also important in the argument that the redundant information provided by multiple statements is the information carried by the truth of their disjunction. One of the two motivations for this idea was based on the principle that the information provided by a statement A is always *part of* the information provided by any stronger statement B. This does not mean however, that statement A necessarily provides *quantitatively* less information than B (i.e. *less bits* of information). In fact, this latter principle would contradict classical information theory. Here is why: suppose the pointwise mutual

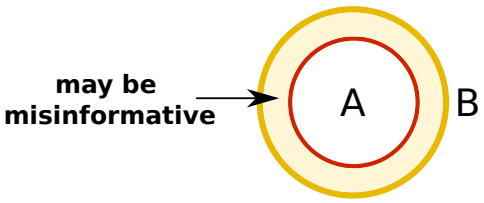

Fig. 4.10: Illustration of the idea that the information provided by a logically weaker statement A is always *part of* the information of a stronger statement B, even though the latter may provide *less bits* of information. This phenomenon can be explained in terms of the misinformative, i.e. negative, contribution of the surplus information provided by B (the shaded ring).

information $i(t:s) = i(t:S=s)$ is negative. Now, consider any *tautology* such as $S = s \vee \neg(S = s)$. Certainly, this statement is *logically weaker* than $S = s$ because a tautology is implied by any other statements. Furthermore, the probability of the tautology being true is equal to 1. Therefore, the information $i(t:S=s \vee \neg(S=s))$ provided by it is equal to 0. But this means $i(t:S=s) < i(t:S=s \vee \neg(S=s))$ even though $S = s \vee \neg(S = s) \dashv S = s$.

Nonetheless, there certainly is a sense in which a stronger statement B provides "more" information than a weaker statement A: the information provided by A is *part of* the information provided by B. If we know B is true than we can by assumption infer that A is true, and hence, we have access to all the information provided by A. The fact that the stronger statement B may nonetheless provide less bits of information can be explained in terms of misinformation: If we know B is true, then we obtain all the information carried by A *plus some additional information*. If it happens that this surplus information is misinformative, i.e. negative, then quantitatively B will provide less information than A. This idea is illustrated in Figure 4.10.

Importantly, the possible negativity and non-monotonicity of $i_\cap^{\text{sx}}$ as well as the potential negativity of π^{sx} can be *completely* explained in terms of misinformative contributions in the following sense: it is possible [93] to uniquely separate $i_\cap^{\text{sx}}$ into an informative part $i_\cap^{\text{sx}+}$ and a misinformative part $i_\cap^{\text{sx}-}$ such that

$$i_\cap^{\text{sx}}(t:\alpha) = i_\cap^{\text{sx}+}(t:\alpha) - i_\cap^{\text{sx}-}(t:\alpha) \tag{4.37}$$

Now, each of these components can be shown to be non-negative and monotonically increasing over the lattice. Moreover, the induced informative and misinformative atoms $\pi^{\text{sx}+}$ and $\pi^{\text{sx}-}$ are non-negative as well [115]. In other words, once we seperate out informative and misinformative components any violations of non-negativity

and monotonicity disappear. Hence, these violations can be fully accounted for in terms of misinformative contributions.

4.7 Conclusion

In this paper we connected PID theory with ideas from mereology, i.e. the study of parthood relations, and formal logic. The main insights derived from these ideas are that the general structure of information decomposition as originally introduced by Williams and Beer [19] can be derived entirely from 1) parthood relations between information contributions and 2) in terms of a hierarchy of logical constraints on how information about the target can be accessed. In this way the theory is set up from the perspective of the atoms of information, i.e. the quantities we are ultimately interested in. The n-sources PID problem has conventionally been approached by defining a measure of redundant information which in turn implies a unique solution for the atoms of information. We showed how such a measure can be defined in terms of the information provided by logical statements of a specific form. We discussed furthermore how the parthood perspective can be utilized to systematically address the question of whether a PID may be determined based on concepts other than redundancy. In doing so, we showed that this is indeed possible in terms of measures of "restricted information", "weak synergy", and "moderate synergy" but not in terms of "strong synergy". We hope to have shown that there are deep connections between mereology, formal logic and information decomposition that future research in these fields may benefit from.

4.8 Appendix

4.8.1 Minimally Consistent PID

Definition 2 (Minimally consistent PID)**.** *Let $S_1, \ldots, S_n, T$ be jointly distributed random variables with joint distribution $\mathbb{P}_J$ and let $\mathcal{B}_n$ be the set of parthood distributions in the context of n source variables. A minimally consistent partial-information-decomposition of the mutual information provided by the sources $S_1, \ldots, S_n$ about the target T is any function $\Pi_{\mathbb{P}_j} : \mathcal{B}_n \to \mathbb{R}$, determined by $\mathbb{P}_J$, that satisfies*

$$I_{\mathbb{P}_J}(T : (S_i)_{i\in\mathbf{a}}) = \sum_{f(\mathbf{a})=1} \Pi_{\mathbb{P}_J}(f) \tag{4.38}$$

for all $\mathbf{a} \subseteq \{1, \ldots, n\}$. *The subscripts* $\mathbb{P}_J$ *indicate that both the mutual information and the information atoms are functions of the underlying joint distribution.*

4.8.2 Proof of isomorphism between $(\mathcal{B}, \sqsubseteq)$, $(\mathcal{L}, \dashv)$ and $(\mathcal{A}, \preceq)$

First, recall that the relation $\models$ of logical implication is formally defined in terms of the notion of a *valuation* [116]. A valuation is an assignment of truth-values (0 for false and 1 for true) over the propositional variables φ_i. So the set of all valuations $\mathcal{V}$ is given by the set of all mappings from $\{\varphi_1, \ldots, \varphi_n\}$ into $\{0, 1\}$:

$$\mathcal{V} := \{0, 1\}^{\{\varphi_1, \ldots, \varphi_n\}} \tag{4.39}$$

A valuation is said to *satisfy* a statement $\tilde{\alpha}$, written as $\models_V \tilde{\alpha}$, under the following conditions

1. If $\tilde{\alpha}$ is an atomic statement, then $\models_V \tilde{\alpha} \iff V(\tilde{\alpha}) = 1$
2. If $\tilde{\alpha}$ is of the form $\tilde{\beta} \wedge \tilde{\gamma}$, then $\models_V \tilde{\alpha} \iff \models_V \tilde{\beta}$ and $\models_V \tilde{\gamma}$
3. If $\tilde{\alpha}$ is of the form $\tilde{\beta} \vee \tilde{\gamma}$, then $\models_V \tilde{\alpha} \iff \models_V \tilde{\beta}$ or $\models_V \tilde{\gamma}$

In this way, the satisfaction relationship is inductively defined for all statements of the propositional language we are considering here. The relation of logical implication is now defined such that a statement $\tilde{\alpha}$ implies a statement $\tilde{\beta}$ just in case all valuations that satisfy $\tilde{\alpha}$ also satisfy $\tilde{\beta}$. Formally,

$$\tilde{\alpha} \models \tilde{\beta} \iff \forall V \in \mathcal{V} : \models_V \tilde{\alpha} \rightarrow \models_V \tilde{\beta} \tag{4.40}$$

Proof of the theorem. We first show the isomorphism between $(\mathcal{B}, \sqsubseteq)$ and $(\mathcal{A}, \preceq)$ and then the isomorphism between $(\mathcal{A}, \preceq)$ and $(\mathcal{L}, \dashv)$. The following mapping $\varphi : \mathcal{A} \to \mathcal{B}$ is an isomorphism between $(\mathcal{B}, \sqsubseteq)$ and $(\mathcal{A}, \preceq)$:

$$\varphi(\alpha) := f_\alpha \text{ with } f_\alpha(\mathbf{b}) = \begin{cases} 1 \text{ if } \exists \mathbf{a} \in \alpha : \mathbf{b} \supseteq \mathbf{a} \\ 0 \text{ otherwise} \end{cases} \tag{4.41}$$

First, φ is surjective: let $f \in \mathcal{B}$, then $\varphi(\alpha_f) = f$ for the set α_f of minimal elements with value 1, i.e.

$$\alpha_f := \{\mathbf{a} \mid f(\mathbf{a}) = 1 \;\&\; \neg \exists \mathbf{b} \subset \mathbf{a} : f(\mathbf{b}) = 1\} \tag{4.42}$$

φ is also injective: let $\varphi(\alpha) = f_\alpha = f_\beta = \varphi(\beta)$ and let $\mathbf{b} \in \beta$. Then, $f_\beta(\mathbf{b}) = 1$ and hence $f_\alpha(\mathbf{b}) = 1$. Therefore, $\exists \mathbf{a} \in \alpha : \mathbf{b} \supseteq \mathbf{a}$. But this can only be true if $\mathbf{b} = \mathbf{a}$, because suppose $\mathbf{b} \supset \mathbf{a}$. We have $f_\beta(\mathbf{a}) = 1$ and hence $\exists \mathbf{b}^* \in \beta : \mathbf{a} \supseteq \mathbf{b}^*$. But then $\mathbf{b} \supset \mathbf{a} \supseteq \mathbf{b}^*$ while $\mathbf{b}, \mathbf{b}^* \in \beta$ contradicting the fact that β is an antichain. Hence, $\mathbf{b} \in \alpha$. By the same argument it can be shown that any $\mathbf{a} \in \alpha$ has to be in β and therefore $\alpha = \beta$.

It remains to be shown that φ is structure preserving. So let $\alpha \preceq \beta$, i.e. $\forall \mathbf{b} \in \beta \exists \mathbf{a} \in \alpha : \mathbf{b} \supseteq \mathbf{a}$. We need to show that in this case $\varphi(\alpha) \sqsubseteq \varphi(\beta)$, i.e. $f_\beta(\mathbf{a}) = 1 \rightarrow f_\alpha(\mathbf{a}) = 1$. So let $f_\beta(\mathbf{a}) = 1$, then $\exists \mathbf{b} \in \beta : \mathbf{a} \supseteq \mathbf{b}$. By assumption this means that $\exists \mathbf{a}^* \in \alpha : \mathbf{b} \supseteq \mathbf{a}^*$. Hence $\mathbf{a} \supseteq \mathbf{a}^*$ and therefore $f_\alpha(\mathbf{a}) = 1$. Regarding the other direction suppose that $f \sqsubseteq g$. Now let $\mathbf{b} \in \beta_g = \varphi^{-1}(g)$, then $g(\mathbf{b}) = 1$ and hence $f(\mathbf{b}) = 1$. Therefore, $\exists \mathbf{a} \in \alpha_f = \varphi^{-1}(f) : \mathbf{b} \supseteq \mathbf{a}$, and thus, $\alpha_f \preceq \beta_g$.

We now turn to the isomorphism between $(\mathcal{L}, \models)$ and $(\mathcal{A}, \preceq)$. The mapping $\Psi : \mathcal{A} \rightarrow \mathcal{L}$ defined in the main text is an isomorphism. Ψ is injective for let $\alpha, \beta \in \mathcal{A}$ be two distinct antichains. Then there has to be an $\mathbf{a} \in \alpha$ not contained in β (or vice versa). But then the conjunction $\bigwedge_{i \in \mathbf{a}} \varphi_i$ will appear in $\tilde{\alpha}$ while it does not appear in $\tilde{\beta}$. Accordingly, $\tilde{\alpha}$ and $\tilde{\beta}$ are distinct elements of $\mathcal{L}$. Ψ is surjective as well for let $\tilde{\alpha} \in \mathcal{L}$. Then $\tilde{\alpha}$ is of the form $\bigvee_{\mathbf{j} \in \mathbf{J}} \bigwedge_{i \in \mathbf{j}} \varphi_i$ for some set of index sets $\mathbf{J} = \{\mathbf{j}_1, \ldots, \mathbf{j}_m\}$ where $\mathbf{j}_i \subseteq \{1, \ldots, n\}$. Because the conjunctions $\bigwedge_{i \in \mathbf{j}} \varphi_i$ have to be logically independent it follows that the index sets cannot be subsets of each other, i.e. $\neg(\mathbf{j}_k \supseteq \mathbf{j}_l)$ for $k \neq l$. But this implies that $\mathbf{J}$ is an antichain which is, by definition of Ψ, mapped onto $\tilde{\alpha}$.

It only remains to be shown that $\beta \preceq \alpha \iff \tilde{\beta} \models \tilde{\alpha}$. First, suppose that $\beta \preceq \alpha$. We need to show that for all valuations $V \in \mathcal{V} = \{0,1\}^{\{\varphi_1, \ldots, \varphi_n\}}$: $\models_V \tilde{\alpha} \rightarrow \models_V \tilde{\beta}$, i.e. all Boolean valuations of the φ_i that make $\tilde{\alpha}$ true, also make $\tilde{\beta}$ true. So suppose $\models_V \tilde{\alpha}$, then there must be an $a \in \alpha$ such that $\models_V \bigwedge_{i \in a} \varphi_i$. But since $\beta \preceq \alpha$, there must be a $\mathbf{b} \in \beta$ such that $a \supseteq b$. Therefore, $\models_V \bigwedge_{i \in b} \varphi_i$. Hence, V also satisfies the disjunction over all $\mathbf{b} \in \beta$: $\models_V \bigvee_{\mathbf{b} \in \beta} \bigwedge_{i \in b} \varphi_i = \tilde{\beta}$.

Regarding the other direction, suppose that $\tilde{\beta} \models \tilde{\alpha}$, i.e. all valutions satisfying $\tilde{\alpha}$ also satisfy $\tilde{\beta}$. Now suppose for contradiction that $\neg(\beta \preceq \alpha)$, i.e. $\exists \mathbf{a}^* \in \alpha \forall \mathbf{b} \in \beta : \neg(\mathbf{a} \supseteq \mathbf{b})$. In this case, we can construct a valuation V that satisfies $\tilde{\alpha}$ but not $\tilde{\beta}$ in the following way:

$$V(\varphi_i) = \begin{cases} 1 & \text{if } i \in \mathbf{a}^* \\ 0 & \text{if } i \notin \mathbf{a}^* \end{cases} \tag{4.43}$$

By construction all $b \in \beta$ contain at least one index i not contained in a. Therefore, V does not satisfy any of the conjunctions $\bigwedge_{i \in b} \varphi_i$, and thus it does not satisfy $\tilde{\beta}$, in contradiction to the initial assumption. Hence, $\beta \preceq \alpha$, concluding the proof. □

Corollary 2. *$(\mathcal{L}, \dashv)$ and $(\mathcal{B}, \sqsubseteq)$ are lattices.*

Proof. Follows from the isomorphism and the fact that $(\mathcal{A}, \preceq)$ is a lattice as shown by [114]. □

4.8.3 Proofs of Propositions

Monotonic truth tables

Proof of Proposition 1. Let $\tilde{\alpha} \in \mathcal{L}$ and let $V, V' \in \mathcal{V}$ such that $\forall i \in \{1, \ldots, n\} : V(\varphi_i) = 1 \rightarrow V'(\varphi_i) = 1$. Suppose that $T_{\tilde{\alpha}}(V) = 1$. Then V must satisfy at least one of the conjunctions $\bigwedge_{i \in \mathbf{a}} \varphi_i$. But since $V(\varphi_i) = 1 \rightarrow V'(\varphi_i) = 1$ any conjunction satisfied by V must also be satisfied by V'. Hence, $T_{\tilde{\alpha}}(V') = 1$.

Regarding the converse: let T be a monotonic truth-table. Then $T = T_{\tilde{\alpha}^*}$ for the statement

$$\tilde{\alpha}^* = \bigvee_{\substack{V \in \mathcal{V} \\ T(V)=1}} \bigwedge_{\substack{i \in \{1,\ldots,n\} \\ V(\varphi_i)=1}} \varphi_i \tag{4.44}$$

Note that $\tilde{\alpha}^*$ is generally not in $\mathcal{L}$ because the conjunctions are not necessarily logically independent. But one can obtain an equivalent statement $\tilde{\alpha} \in \mathcal{L}$ by removing all conjunctions from $\tilde{\alpha}^*$ that logically imply another conjunction in $\tilde{\alpha}^*$. Let $\tilde{\alpha}$ be this statement. Then, if $\tilde{\alpha}$ is true, certainly $\tilde{\alpha}^*$ is true because the latter differs from the former only through additional disjuncts. Conversely, if $\tilde{\alpha}^*$ is true, then one of its conjuncts must be true. If the true conjunct in $\tilde{\alpha}^*$ does appear in $\tilde{\alpha}$ as well (i.e. it has not been removed), then trivially $\tilde{\alpha}$ has to be true as well. On the other hand, if this conjunct does not appear in $\tilde{\alpha}$, then it must have been removed which implies that there is a logically weaker conjunct in $\tilde{\alpha}$. But then this logically weaker conjunct has to be true as well, thereby making $\tilde{\alpha}$ true. Therefore, $\tilde{\alpha}^*$ and $\tilde{\alpha}$ have the same truth-table T and $\tilde{\alpha} \in \mathcal{L}$ as desired. Furthermore, $\tilde{\alpha}$ is unique because $\models$ is antisymmetric on $\mathcal{L}$ by Corollary 1. Hence, there can be no two distinct but logically equivalent elements (i.e. elements with the same truth-table) in $\mathcal{L}$. □

Characterization of Children

Proof of Proposition 2. Concerning the if-part we show the contraposition: Suppose that there is a $\tilde{\beta}$ strictly in between $\tilde{\gamma}$ and $\tilde{\alpha}$. If this is the case, then there must be a model V_1 such that $T_{\tilde{\beta}}(V_1) = 1$ while $T_{\tilde{\alpha}}(V_1) = 0$ and a distinct model V_2 such that

$T_{\tilde{\gamma}}(V_2) = 1$ while $T_{\tilde{\beta}}(V_2) = 0$. But for both of these models it would be true that $T_{\tilde{\gamma}}(V_1) = 1$ while $T_{\tilde{\alpha}}(V_1) = 0$. Thus, $\tilde{\gamma}$ would be true in at least two additional cases.

Concerning the only-if part we show the contraposition again: Suppose that $\tilde{\gamma}$ is true in the $k \geqslant 2$ additional cases contained in $\mathcal{V}_* = \{V_1, V_2, \ldots, V_k\}$. Consider the subset of these models with the smallest number of ones:

$$\mathcal{V}_*^{min} = \left\{ V \in \mathcal{V}_* \mid \forall V' \in \mathcal{V}_* : \sum_{i=1}^{n} V(\varphi_i) \leqslant \sum_{i=1}^{n} V'(\varphi_i) \right\} \tag{4.45}$$

Now let $V_* \in \mathcal{V}_*^{min}$. Then the truth table

$$T_{\tilde{\beta}}(V) := \begin{cases} 1 & \text{if } T_{\tilde{\gamma}}(V) = 1 \text{ but } V \neq V_* \\ 0 & \text{otherwise} \end{cases} \tag{4.46}$$

is monotonic and the statement $\tilde{\beta}$ associated with this truth-table is strictly in between $\tilde{\gamma}$ and $\tilde{\alpha}$. The latter is true because all valuations that satisfy $\tilde{\alpha}$ also satisfy $\tilde{\beta}$ and all valuations that satisfy $\tilde{\beta}$ also satisfy $\tilde{\gamma}$. At the same time there is a valuation, namely V_*, that satisfies $\tilde{\gamma}$ but not $\tilde{\beta}$, and a set of valuations with at least one element, namely $\mathcal{V}_* \backslash \{V_*\}$, that satisfies $\tilde{\beta}$ but not $\tilde{\alpha}$. Thus, all three statements have to be distinct. Regarding the monotonicity: by assumption $\tilde{\gamma}$ has a monotonic truth-table and the truth-table of $\tilde{\beta}$ is identical except that $T_{\tilde{\beta}}(V_*) = 0$. So the only way $T_{\tilde{\beta}}$ could *not* be monotonic would be for there to exist a valuation V'_*, distinct from V_*, that would enforce $T_{\tilde{\beta}}(V_*) = 1$ via monotonicity, i.e. a valuation that results from flipping some ones in V_* to zeros and that satisfies $\tilde{\beta}$. Suppose there is such a valuation. V'_* would have to satisfy $\tilde{\beta}$ while not satisfying $\tilde{\alpha}$, since if it did satisfy $\tilde{\alpha}$, V_* would have to satisfy $\tilde{\alpha}$ as well in contradiction to $V_* \in \mathcal{V}_*$. Furthermore, as V'_* satisfies $\tilde{\beta}$ it also satisfies $\tilde{\gamma}$. Therefore, $V'_* \in \mathcal{V}_*$. However, if it were true that $V'_*(\varphi_i) = 1 \rightarrow V_*(\varphi_i) = 1$, then $\sum_{i=1}^{n} V'_*(\varphi_i) < \sum_{i=1}^{n} V_*(\varphi_i)$, contradicting the fact that $V_* \in \mathcal{V}_*^{min}$. □

Lower bound on children

Proof of Proposition 3. Let α be such an antichain and let $\mathbf{a} \in \alpha$ be a set of indices such that $|\mathbf{a}| = k$. We utilize the isomorphism between $\mathcal{A}$ and $\mathcal{L}$ by showing that $\tilde{\alpha}$

has at least k children. Since $|\mathbf{a}| = k$ there are exactly k distinct indices $i_1, \ldots, i_k \in \mathbf{a}$ and we can define k subsets of valuations

$$\mathcal{V}_1 = \{V \in \mathcal{V} : \neg(\models_V \tilde{\alpha}) \;\&\; i \in \mathbf{a} \backslash \{i_1\} \rightarrow V(\varphi_i) = 1\} \tag{4.47}$$

$$\ldots$$

$$\mathcal{V}_k = \{V \in \mathcal{V} : \neg(\models_V \tilde{\alpha}) \;\&\; i \in \mathbf{a} \backslash \{i_k\} \rightarrow V(\varphi_i) = 1\} \tag{4.48}$$

In other words, the valuations in $\mathcal{V}_1$, first, do not satisfy $\tilde{\alpha}$, and second, assign a one to all φ_i if i is in the collection $\mathbf{a}$ but not equal to i_1. The definition of the other $\mathcal{V}_i$ is analogous. The goal is now to find 'maximal' valuations (making as many φ_i true as possible) in these sets and modify the truth-table of $\tilde{\alpha}$ by assigning a one to exactly one of these valuations. This can be done for all valuations separately to obtain k novel monotonic truth-tables. These monotonic truth-tables are uniquely associated with specific statements via Proposition 1 which can then be shown to be children by Proposition 2 since they are true in exactly one more case than $\tilde{\alpha}$. To make this argument note first that $\mathcal{V}_1, \ldots, \mathcal{V}_k$ each contain at least one element $V_1, \ldots, V_k$ respectively:

$$V_1(\varphi_i) = \begin{cases} 1 & \text{if } i \in \mathbf{a} \backslash \{i_1\} \\ 0 & \text{otherwise} \end{cases} \tag{4.49}$$

$$\ldots$$

$$V_k(\varphi_i) = \begin{cases} 1 & \text{if } i \in \mathbf{a} \backslash \{i_k\} \\ 0 & \text{otherwise} \end{cases} \tag{4.50}$$

These valuations do not satisfy $\tilde{\alpha}$: They don't satisfy the conjunction $\bigwedge_{i \in \mathbf{a}} \varphi_i$ and since α is an antichain each $\mathbf{a}' \neq \mathbf{a}$ has to contain at least one index j not contained in $\mathbf{a}$. The corresponding conjunctions $\bigwedge_{i \in \mathbf{a}'} \varphi_i = \varphi_j \wedge \bigwedge_{i \in \mathbf{a}' \backslash \{j\}} \varphi_i$ are therefore not satisfied by any V_i since by construction $V_1(\varphi_j) = \ldots = V_k(\varphi_j) = 0$. Now consider the sets of 'maximal' valuations within the $\mathcal{V}_i$:

$$\mathcal{V}_1^{max} = \left\{ V \in \mathcal{V}_1 | \forall V' \in \mathcal{V}_1 : \sum_{i=1}^{n} V'(\varphi_i) \leqslant \sum_{i=1}^{n} V(\varphi_i) \right\} \tag{4.51}$$

$$\ldots$$

$$\mathcal{V}_k^{max} = \left\{ V \in \mathcal{V}_k | \forall V' \in \mathcal{V}_k : \sum_{i=1}^{n} V'(\varphi_i) \leqslant \sum_{i=1}^{n} V(\varphi_i) \right\} \tag{4.52}$$

Let $V_1^* \in \mathcal{V}_1^{max}, ..., V_k^* \in \mathcal{V}_k^{max}$. Due to the maximality of these valuations the following truth-tables are monotonic

$$T_{\tilde{\gamma}_1}(V) = \begin{cases} 1 & \text{if } T_{\tilde{\alpha}}(V) = 1 \text{ or } V = V_1^* \\ 0 & \text{otherwise} \end{cases} \tag{4.53}$$

$$\dots$$

$$T_{\tilde{\gamma}_k}(V) = \begin{cases} 1 & \text{if } T_{\tilde{\alpha}}(V) = 1 \text{ or } V = V_k^* \\ 0 & \text{otherwise} \end{cases} \tag{4.54}$$

This is because, first, the truth-table of $\tilde{\alpha}$ is already monotonic, and second, if a zero is flipped to a one in V_1^* or $\dots$ or M_k^* the resulting valuations are by construction guaranteed to satisfy $\tilde{\alpha}$. Otherwise, we would obtain valuations in $\mathcal{V}_1$ or $\dots$ or $\mathcal{V}_k$ respectively, containing more ones than V_1^* or $\dots$ or V_k^* respectively, in contradiction to the maximality of these valuations. The uniquely defined statements $\tilde{\gamma}_1, \dots, \tilde{\gamma}_k$ corresponding to these truth-tables via Proposition 1 are children of $\tilde{\alpha}$ by Proposition 2 because each of them is true in exactly one additional valuation compared to $\tilde{\alpha}$. Finally all of these statements are distinct since they are pairwise logically independent and a single statement cannot have multiple truth-tables. □

Algorithm to determine children

Proof of Proposition 4. Firstly, any $\tilde{\gamma}$ produced by the algorithm is a direct child since its truth-table differs from that of $\tilde{\alpha}$ only through an additional one, i.e. $\tilde{\gamma}$ is true in exactly one more case than $\tilde{\alpha}$ and is thus a direct child by Proposition 2. Secondly, there is no child of $\tilde{\alpha}$ that is not generated by the algorithm. Again by Proposition 2, the truth-table of any such child would differ from that of $\tilde{\alpha}$ only through a single one. But the algorithm explores systematically *all* possibilities to add a single one to the truth-table of $\tilde{\alpha}$. Thus any child $\tilde{\gamma}$ will be generated at some point. □

A pseudocode version of the algorithm is shown in Algorithm 1.

Meet and Join operations on logic lattices

The meet $\tilde{\wedge}$ and join $\tilde{\vee}$ operations can be explicitly constructed in the following way: The element of $\mathcal{L}$ logically equivalent to the disjunction $\tilde{\alpha} \vee \tilde{\beta}$ can be obtained by simply removing all disjuncts that logically imply another disjunct. The element of $\mathcal{L}$ logically equivalent to the conjunction $\tilde{\alpha} \wedge \tilde{\beta}$ can be obtained by, first, applying

Algorithm 1: Determines children of a statement $\tilde{\alpha}$ in the logic lattice.

GetChld $\tilde{\alpha}$
  inputs : A statement $\tilde{\alpha}$
  outputs : The set of children of $\tilde{\alpha}$ denoted by $\mathcal{C}_{\tilde{\alpha}}$
  $k \leftarrow 0$
  $\mathcal{V}_{\tilde{\alpha}} \leftarrow \varnothing$
  $\mathcal{C}_{\tilde{\alpha}} \leftarrow \varnothing$
  `// step (1)`
  foreach *valuation* $V \in \mathcal{V}$ **do**
    if $\not\models_V \tilde{\alpha}$ **then**
      $\mathcal{V}_{\tilde{\alpha}} \leftarrow \mathcal{V}_{\tilde{\alpha}} \cup V$
      `// Maximal number of ones in` V `if` $\not\models_V \tilde{\alpha}$
      if $\sum_i V_i > k$ **then**
        $k \leftarrow \sum_i V_i$
  `// step (3) as a while loop`
  while $k \neq 0$ **do**
    `// Construct the set of all` $V \in \mathcal{V}_{\tilde{\alpha}}$ `such that` $\sum_i V_i = k$
    $\mathcal{V}_{\tilde{\alpha}}^k \leftarrow \varnothing$
    foreach *valuation* $V \in \mathcal{V}_{\tilde{\alpha}}$ **do**
      if $\sum_i V_i = k$ **then**
        $\mathcal{V}_{\tilde{\alpha}}^k \leftarrow V$
    `// Construct a child of` $\tilde{\alpha}$ `if it exists (step (2))`
    foreach *valuation* $V \in \mathcal{V}_{\tilde{\alpha}}^k$ **do**
      $Q \leftarrow \varnothing$
      for $V' \in \mathcal{V}_{\tilde{\alpha}}$ **do**
        if $\sum_i V'_i = k+1$ *and* $V(\varphi_i) = 1 \rightarrow V'(\varphi_i) = 1 \; \forall i \in [n]$ **then**
          $Q \leftarrow V'$
          break
      if $Q = \varnothing$ **then**
        construct $\tilde{\gamma}$ that satisfies V and every $V' \in \mathcal{V} \backslash \mathcal{V}_{\tilde{\alpha}}$
        $\mathcal{C}_{\tilde{\alpha}} \leftarrow \mathcal{C}_{\tilde{\alpha}} \cup \tilde{\beta}$
    $k \leftarrow k - 1$
  return $\mathcal{C}_{\tilde{\alpha}}$

the distributive law to obtain a disjunction of conjunctions, second, applying the idempotency law to all conjunctions to remove repeated statements, and third, removing again all disjuncts that logically imply another disjunct. Denoting these three operations by $\mathcal{D}$, $\mathcal{I}$, and $\underline{\circ}$ (underline) respectively, the meet and join have the explicit expressions given in the following proposition:

Proposition 12 (Meet and Join Operations)**.**

$$\tilde{\alpha} \tilde{\wedge} \tilde{\beta} = \underline{\tilde{\alpha} \vee \tilde{\beta}} \tag{4.55}$$

$$\tilde{\alpha} \tilde{\vee} \tilde{\beta} = \underline{\mathcal{I}(\mathcal{D}(\tilde{\alpha} \wedge \tilde{\beta}))} \tag{4.56}$$

Proof. By construction, $\underline{\tilde{\alpha} \vee \tilde{\beta}}$ and $\underline{\mathcal{I}(\mathcal{D}(\tilde{\alpha} \wedge \tilde{\beta}))}$ are in $\mathcal{L}$. Furthermore, since the operations $\mathcal{D}$, $\mathcal{I}$, and $\underline{\circ}$ do not affect the truth-conditions of statements, $\underline{\tilde{\alpha} \vee \tilde{\beta}}$ and $\underline{\mathcal{I}(\mathcal{D}(\tilde{\alpha} \wedge \tilde{\beta}))}$ are logically equivalent to $\tilde{\alpha} \vee \tilde{\beta}$ and $\tilde{\alpha} \wedge \tilde{\beta}$, respectively. Hence, it only needs to be shown that these latter statements satisfy the conditions of meet and join respectively. Now, clearly $\tilde{\alpha} \vee \tilde{\beta}$ is logically weaker than both $\tilde{\alpha}$ and $\tilde{\beta}$ while $\tilde{\alpha} \wedge \tilde{\beta}$ is logically stronger than both $\tilde{\alpha}$ and $\tilde{\beta}$. It remains to be shown that former is the strongest such statement while the latter is the weakest such statement. Suppose there was statement $\tilde{\gamma}$ stronger than $\tilde{\alpha} \vee \tilde{\beta}$, then there would have to be a model M^* making $\tilde{\gamma}$ false and $\underline{\tilde{\alpha} \vee \tilde{\beta}}$ true. But since $\underline{\tilde{\alpha} \vee \tilde{\beta}}$ is true whenever either $\tilde{\alpha}$ is true or $\tilde{\beta}$ is true, this means that $\tilde{\gamma}$ would have to be false in a case where one of $\tilde{\alpha}$ or $\tilde{\beta}$ is true. However, this implies that $\tilde{\gamma}$ cannot be logically weaker than both $\tilde{\alpha}$ and $\tilde{\beta}$, and hence, $\underline{\tilde{\alpha} \vee \tilde{\beta}}$ must be the strongest statement logically weaker than $\tilde{\alpha}$ and $\tilde{\beta}$. Now suppose there was a statement $\tilde{\gamma}$ weaker than $\tilde{\alpha} \wedge \tilde{\beta}$, then there would have to be a model M^* making $\tilde{\gamma}$ true but $\tilde{\alpha} \wedge \tilde{\beta}$ false. But this means that $\tilde{\gamma}$ would be true in a case in which either $\tilde{\alpha}$ or $\tilde{\beta}$ is false. Accordingly, $\tilde{\gamma}$ cannot be stronger than both $\tilde{\alpha}$ and $\tilde{\beta}$, and hence, $\underline{\mathcal{I}(\mathcal{D}(\tilde{\alpha} \wedge \tilde{\beta}))}$ must be the weakest statement logically stronger than $\tilde{\alpha}$ and $\tilde{\beta}$. □

4.8.4 Derivations related to restricted information based and synergy based PID

Relation between restricted information and conditional mutual information

The relation between restricted information and conditional mutual information given by Equation 5.5 can be derived via the chain rule as follows:

$$
\begin{aligned}
I\left(T:(S_i)_{i\in\alpha_\cup}|(S_j)_{j\in\alpha_\cup^C}\right) &= I(T:(S_i)_{i\in[n]}) - I(T:(S_j)_{j\in\alpha_\cup^C}) && (4.57)\\
&= \sum_{f([n])=1}\Pi(f) - \sum_{f(\alpha_\cup^C)=1}\Pi(f) && (4.58)\\
&= \sum_{f(\alpha_\cup^C)=0}\Pi(f) && (4.59)\\
&= \sum_{f(\mathbf{b})=1\to\exists j:\{i_j\}\supseteq\mathbf{b}}\Pi(f) && (4.60)\\
&= I_{\text{res}}(T:\alpha) && (4.61)
\end{aligned}
$$

Proof that moderate synergy induces a unique PID

The claim that defining a measure of moderate synergy leads to a unique solution for the atoms of information can be shown by starting from the system of equation associated with weak synergy. These equations can be transformed into the moderate synergy equations by operations that preserve invertibility. First, the "self-synergy" equations

$$
I_{\text{ws/ms}}(T:\mathbf{a}) = I(T:\mathbf{a}^C|\mathbf{a}) = \sum_{f(\mathbf{a})=0}\Pi(f) \tag{4.62}
$$

are contained in both systems. Furthermore, weak and moderate synergy coincide if $\alpha_\cup = [n]$. In this case, the additional constraint $f(\alpha_\cup) = 1$ is superfluous since $f([n])$ is necessarily equal to 1 by the properties of parthood distributions. Thus, the corresponding equations are again contained in both systems. This only leaves the case of $\alpha_\cup \subset [n]$ while $|\alpha| \geqslant 2$. Let α be such an antichain. It can be shown that the corresponding moderate synergies can be expressed as a difference between two equations in the weak synergy system:

$$
\begin{aligned}
I_{\text{ws}}(T:\alpha) - I_{\text{ms}}(T:\alpha) &= \sum_{\substack{\forall\mathbf{a}_i:f(\mathbf{a}_i)=0\\ f(\alpha_\cup)=0}}\Pi(f) && (4.63)\\
&= \sum_{f(\alpha_\cup)=0}\Pi(f) && (4.64)\\
&= I(T:\alpha_\cup^C|\alpha_\cup) && (4.65)
\end{aligned}
$$

where the second to last equality follows because the monotonicity of parthood distributions implies that $f(\alpha_\cup) = 0 \rightarrow f(\mathbf{a}) = 0 \; \forall \mathbf{a} \in \alpha$. Therefore, we obtain

$$I_{\text{ms}}(T : \alpha) = I_{\text{ws}}(T : \alpha) - I(T : \alpha^C_\cup | \alpha_\cup) \tag{4.66}$$

$$= I_{\text{ws}}(T : \alpha) - I_{\text{ws}}(T : \alpha_\cup) \tag{4.67}$$

showing that the moderate synergy equation associated with α is the difference between two weak synergy equations. Since subtracting two equations from each other leaves invertibility unaffected this establishes that the moderate synergy system of equations is invertible as well.

4.9 Acknowledgements

MW, AM, AG are employed at the Campus Institute for Dynamics of Biological Networks (CIDBN) funded by the Volkswagen Stiftung. MW, AM received support from the Volkswagenstiftung under the programme 'Big Data in den Lebenswissenschaften'. This work was supported by a funding from the Ministry for Science and Education of Lower Saxony and the Volkswagen Foundation through the "Niedersächsisches Vorab". We thank Kyle Schick-Poland, David Ehrlich, and Andreas Schneider for helpful comments on the draft.

4.10 Author contributions

AG conceived the parthood-based and logic-based formulations of PID and wrote the original manuscript except for the introduction which was provided by MW. MW originally conceived the $i^{sx}_\cap$ measure of redundant information rederived in §4.3. AM provided critical feedback regarding the mathematical aspects of the paper. All authors were involved in revising the manuscript and refining the ideas presented therein. All authors gave final approval for publication and agree to be held accountable for the work performed therein.

5

From Babel to Boole: The Logical Organization of Information Decompositions

Aaron J. Gutknecht [1*] Abdullah Makkeh [1], Michael Wibral [1]

1 Campus Institute for Dynamics of Biological Networks, Georg-August University, Goettingen, Germany

Preprint available at: ***Gutknecht, A. J., Makkeh, A., & Wibral, M. (2023). From Babel to Boole: The Logical Organization of Information Decompositions. arXiv preprint arXiv:2306.00734.***

Abstract

The conventional approach to the general Partial Information Decomposition (PID) problem has been redundancy-based: specifying a measure of redundant information between collections of source variables induces a PID via Moebius-Inversion over the so called redundancy lattice. Despite the prevalence of this method, there has been ongoing interest in examining the problem through the lens of different base-concepts of information, such as synergy, unique information, or union information. Yet, a comprehensive understanding of the logical organization of these different based-concepts and their associated PIDs remains elusive. In this work, we apply the mereological formulation of PID that we introduced in a recent paper to shed light on this problem. Within the mereological approach base-concepts can be expressed in terms of conditions phrased in formal logic on the specific parthood relations between the PID components and the different mutual information terms. We set forth a general pattern of these logical conditions of which all PID base-concepts in the literature are special cases and that also reveals novel base-concepts, in particular a concept we call “vulnerable information”.

5.1 Introduction

Partial information decomposition (PID) is a powerful framework for dissecting the intricate relationships among multiple information sources and their joint contributions to a target variable. In the most simple case of two source variables S_1 and S_2 there is general agreement that the decomposition should contain four terms: the redundant information of S_1 and S_2 about T, the unique information of S_1 about T, the unique information of S_2 about T, and the synergistic information of S_1 and S_2 about T. There is also general agreement that these components should be related to the mutual information provided by subsets of these sources via the equations

$$I(S_1, S_2 : T) = R + U_1 + U_2 + S \tag{5.1}$$

$$I(S_1 : T) = R + U_1 \tag{5.2}$$

$$I(S_2 : T) = R + U_2 \, . \tag{5.3}$$

This system does not have a unique solution for the four components because we are short of one equation. A widely used approach to arrive at a determinate information decomposition is to fix one of the components and solve for the others using the above equation. The component most widely used for this purpose is the redundancy [19, 91, 96, 115, 120–127]. But there are also some unique-information-based [83, 128, 129] and some synergy-based approaches [119, 130]. In principle, it is also possible to fix not an individual component but a certain combination of them, if this combination has an intuitive meaning. An example for this is the sum of the redundancy atom and the two unique information atoms. This describes the entirety of the information we can get from at least one information source and has been called *union information*. [131] and [127] used this as the starting point to fix an information decomposition. We will refer to the information quantity fixed in order to determine a full information decomposition as a PID *base-concept*. In the general n-sources case, a base-concept will in fact encompass a whole set of quantities because the underlying system of equations becomes more and more undetermined. The key objective of this paper is provide a systematic study of PID base-concepts in this general case utilizing the mereological approach to partial information decomposition we introduced in [132].

There are three important reasons why this issue is of interest: First, there is a theoretical reason. Knowledge about the different possible ways to induce a PID provides insights into the structure of the problem. It makes clear which aspects of the original exposition of PID theory are essential and which aspects are replaceable. For example, does the concept of redundancy have a privileged role in PID theory?

What about its underlying lattice structure? Furthermore, addressing the problem from the perspective of other base-concepts may also lead to constraints on possible solutions. For example, there have been numerous proposals for desirable properties or axioms on measures of redundant information [123, 133], and also some for properties of synergistic information [83]. A full account of synergy-based PID establishes a numerical connection between redundancy and synergy that allows us to determine whether the proposed properties of these two concepts are compatible. When they are not, the space of possible solutions to the problem is restricted.

The second reason pertains to the interpretation of information components. While it is true that in principle all base-concepts determine each other (fixing one, fixes all of them), the interpretation of measures that are not used as the base-concept will inevitably be that of a "remainder". Consider the case of two sources: if we specify their redundancy, then we can compute the unique information of each source by subtracting that redundancy from the total information provided by that source about the target. The synergy is then computed by subtracting redundancy and unique information from the total mutual information provided by both sources jointly, i.e. synergy is whatever remains if we subtract the other components from the total. This indirect definition makes the resulting notion of synergy quite intangible. By contrast, in a synergy-based PID, the synergy is directly defined in terms of the underlying joint distribution. This provides us with more control over its interpretation. Of course, the interpretational problem just described is shifted towards the non-synergistic components in this case. However, if in the application at hand synergy is of particular importance, a synergy-based decomposition might be preferable.

The third reason is a computational one. The number of distinct components in a PID grows super-exponentially with the number of information sources. Thus it becomes important to be able to compute useful summaries of the PID that do not require the computation of all components. An example for such a summary is the backbone decomposition introduced by [119]. The components in this decomposition measure the information about the target for which access to exactly k sources is required ($k = 1, \ldots, n$). In this way the components provide a useful measure of the k-way interaction within the system of sources. The backbone components can be calculated very easily from a measure of synergy whereas it is not known how to compute them from a redundancy measure without having to compute all PID components. The same is true for the measure of "representational complexity" introduced in [134]. Approaching the problem from the perspective of synergistic information makes this measure applicable to far larger networks since the computational cost scales only linearly with the number of sources in this case.

Our approach is as follows: In the next section, we review the mereological approach to PID using the example of redundant information and show how it expresses PID base-concepts in terms of their characteristic logical conditions on parthood relations. In Section 5.3 we apply the approach to the construction of synergy-based PIDs. The analyses of redundancy-based and synergy-based PID naturally suggest a more general logical pattern of conditions for defining based-concepts which we will discuss in Section 5.4. The resulting scheme comprises all base-concepts considered in the literature and also leads to new base-concepts. In particular a quantity we call "vulnerable information" and certain "partner measures" of the existing base-concepts which pick out the same information components but viewed from the perspective of different source collections. Section 5.5 addresses the implied properties of the different base-concepts as well as their associated lattices. Finally, in Section 5.6 we discuss the relation of the ideas presented here to some previous approaches before presenting some general conclusions of our analysis in Section 5.7.

5.2 The mereological approach to PID

In a recent paper we showed how to derive PID theory from considerations of parthood relations between information contributions [132]. The key idea is that PID decomposes the information that the sources carry about the target into atomic contributions characterized by their parthood relations to the information provided by the different possible subsets of source variables. In other words, each information atom quantifies precisely that portion of the joint mutual information that stands in a particular constellation of parthood relationships to the 2^n different $I(\mathbf{a} : T)$ terms. Such constellations can be described by what we call *parthood distributions*, i.e. Boolean functions $f : \mathcal{P}(\{1, \ldots, n\}) \to \{0, 1\}$ that tell us for any subset $\mathbf{a} \subseteq \{1, \ldots, n\}$ of sources (referred to via their indices) whether the information atom described by f is part of $I(\mathbf{a} : T)$. Parthood distributions form the cornerstone of mereological PID. Formally, they are defined as follows

Definition 3 (Parthood Distribution)**.** *A parthood distribution is a function* $f : \mathcal{P}(\{1, \ldots, n\}) \to \{0, 1\}$ *such that*

1. $f(\varnothing) = 0$ *("There is no information in the empty set of sources")*
2. $f(\{1, \ldots, n\}) = 1$ *("All information is in the full set of sources")*
3. $\mathbf{a} \subseteq \mathbf{b}$ & $f(\mathbf{a}) = 1 \to f(\mathbf{b}) = 1$ *("All information in a set of sources is also in all of its supersets")*

In a partial information decomposition there is one information atom $\Pi(f)$ per parthood distribution f. These considerations already tell us the intended meaning of the information atoms and how many atoms there are: one per parthood distribution. Since parthood distributions are formally non-constant, monotonic Boolean functions their number for n sources is equal to the n-th Dedekind number minus two. Now, the question is: how many bits of information does each atom provide? In order to answer this question it is fruitful to think about how the atoms should be related to already known information quantities like mutual information. Given how the atoms are characterized it seems reasonable to demand that the following relation should be satisfied:

$$I(\mathbf{a} : T) = \sum_{f(\mathbf{a})=1} \Pi(f) \quad (\textbf{consistency equation}) \tag{5.4}$$

This equation simply says that any mutual information should made up of all atoms which are part of it. And these are by construction all atoms $\Pi(f)$ such that $f(\mathbf{a}) = 1$. Summing over all such atoms will therefore yield the mutual information carried by the collection $\mathbf{a}$ about the target. We call Equation 5.4 the *consistency equation* of PID. It provides constraints on quantitative solutions for the atoms $\Pi(f)$ by requiring them to be related to mutual information in a particular way.

It is well known, however, that the consistency equation alone still leaves the problem severely under-constrained. We need some additional requirements to obtain a unique solution. This is traditionally achieved by invoking the concept of redundant information, which we generically denote by $I_\cap$. Based on the intended meaning of the atoms we should have

$$I_\cap(\mathbf{a}_1, \ldots, \mathbf{a}_m : T) = \sum_{\forall i f(\mathbf{a}_i)=1} \Pi(f) \tag{5.5}$$

In other words, the information shared by collections $\mathbf{a}_1, \ldots, \mathbf{a}_m$ about T should consist of all information atoms that are part of *each* of the $I(\mathbf{a}_i : T)$ contributions. But these are of course exactly those atoms $\Pi(f)$ such that $f(\mathbf{a}) = 1$ for *all* $i = 1, \ldots, m$. It can be shown that Equation 5.5 is invertable so that once a measure of redundancy is specified a unique solution for the information atoms is implied. A PID obtained in this way is called a *redundancy-based* PID. To see how this works, two insights are crucial.

First, note that Equation 5.5 places a number of constraints on redundancy functions $I_\cap$:

1. $I_\cap(\mathbf{a}_1, \ldots, \mathbf{a}_m : T) = I_\cap(\mathbf{a}_{\sigma(1)}, \ldots, \mathbf{a}_{\sigma(m)} : T)$ for any permutation σ **(symmetry)** (5.6)
2. If $\mathbf{a}_i \supseteq \mathbf{a}_j$ for $i \neq j$, then $I_\cap(\mathbf{a}_1, \ldots, , \mathbf{a}_m : T) = I_\cap(\mathbf{a}_1, \ldots, \mathbf{a}_{i-1}, \mathbf{a}_{i+1}, \ldots, \mathbf{a}_m : T)$ (5.7)
   (superset invariance)
3. $I_\cap(\mathbf{a} : T) = I(\mathbf{a} : T)$ **(self-redundancy)** (5.8)

These constraints follow immediately from the properties of parthood distributions described above. In the literature they are known as the "Williams and Beer axioms" for redundancy functions since in their original exposition these properties play the role of axioms instead of being implied properties. The first two of them, symmetry and superset invariance, imply that the domain of redundancy functions can be reduced to the set of antichains of the partial order $(\mathcal{P}(\{1, \ldots, n\}), \subseteq)$. We use the symbol $\mathcal{A}$ to denote the set of antichains without $\{\}$ and $\{\{\}\}$ since these do not correspond to any meaningful redundancy terms.

The second important idea is that information atoms can be ordered quite naturally according to “how easily” they can be accessed. This can be expressed formally in terms of the following ordering on the parthood distributions:

$$f \sqsubseteq g \Leftrightarrow (\forall\ \mathbf{a} \quad g(\mathbf{a}) = 1 \rightarrow f(\mathbf{a}) = 1) \tag{5.9}$$

Intuitively, whenever the information described by g is accessible via some collection, the information described by f is also accessible via this collection. This order relation constitutes a lattice at the top of which we find the all-way synergy that can only be accessed if we know all sources and at the bottom of which we find the all-way redundancy that can be accessed via any source. Now this ordering stands in a close relationship to the concept of redundant information as expressed in Equation 5.5: consider an antichain $\alpha = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ and the parthood distribution that assigns the value one to *exactly* these collections and their supersets. We denote this distribution by f_α. We know that in general the redundant information $I_\cap(\mathbf{a}_1, \ldots, \mathbf{a}_m : T)$ is equal to the sum of all atoms with parthood distributions assigning the value one to all of the $\mathbf{a}_i$. But these atoms are, by construction, all atoms below and including f_α in the lattice. So we can rewrite Equation 5.5 as

$$I_\cap(\alpha : T) = \sum_{g \sqsubseteq f_\alpha} \Pi(g) \tag{5.10}$$

The mapping $\alpha \to f_\alpha$ also induces a lattice structure on $\mathcal{A}$ that describes the nesting of redundant information terms. The induced ordering is

$$\alpha \preceq \beta \Leftrightarrow f_\alpha \sqsubseteq f_\beta \tag{5.11}$$

The redundant information associated with an antichain α is included in any redundancy associated with antichains β higher up in the lattice. The lattice $(\mathcal{A}, \preceq)$ is the familiar *redundancy lattice* intitally introduced by Williams and Beer [19]. By construction the mapping $\alpha \to f_\alpha$ is an isomorphism between the redundancy lattice and the parthood lattice. The inverse is given by

$$f \to \alpha_f = \{\mathbf{a} | f(\mathbf{a}) = 1 \ \& \ \neg\exists \mathbf{b} \subset \mathbf{a} f(\mathbf{b}) = 1\} \tag{5.12}$$

In other words, α_f is the set of minimal collections (with respect to $\subseteq$) that are assigned the value 1 by f. With these mappings one may also write Equation 5.5 as a Moebius-Inversion over either $\mathcal{A}$ or $\mathcal{B}$ using the conventions $\Pi(\alpha) := \Pi(f_\alpha)$ and $I_\cap(f : T) := I_\cap(\alpha_f : T)$:

$$I_\cap(\alpha : T) = \sum_{\beta \preceq \alpha} \Pi(\beta) \qquad I_\cap(f : T) = \sum_{g \sqsubseteq f} \Pi(g) \tag{5.13}$$

These are now standard Moebius-Inversion formulas which are known to have a unique solution once a measure of redundant information $I_\cap$ is specified. This completes the redundancy-based PID story up to the choice of a concrete redundancy measure. Figure 5.1 illustrates the parthood and redundancy lattices as and how redundancy terms are expressed in terms of information atoms for the case $n = 3$. In the next section, we apply the same mereological ideas in order to address the question of how a synergy-based PID can be constructed.

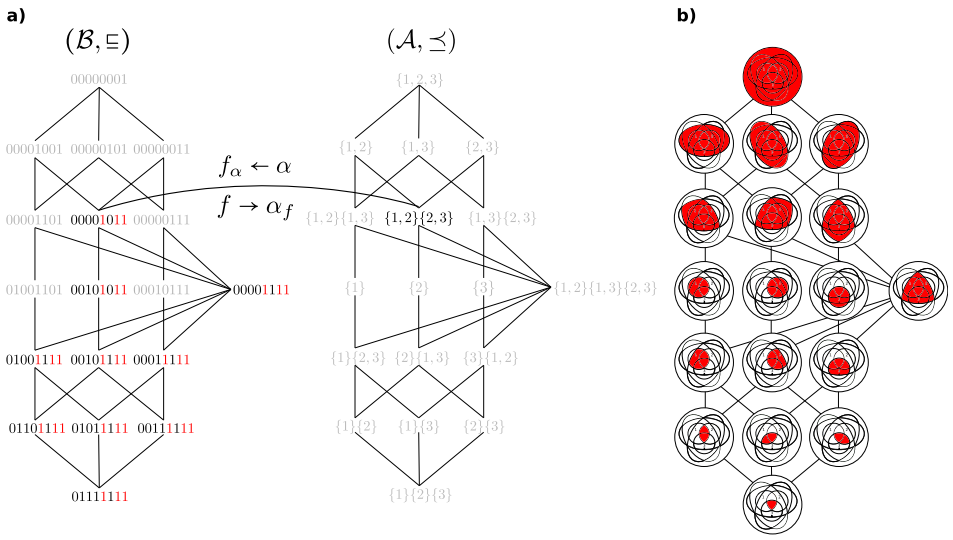

Fig. 5.1: **a)** Parthood and redundancy lattices for $n = 3$ sources. There is an isomorphism between the lattices such that the redundancy associated with a node in the redundancy lattice is equal to the sum of atoms associated with parthood distributions below and including the corresponding node in the parthood lattice. This is shown for the antichain $\{1, 2\}\{2, 3\}$. Note that we adhere to the standard convention of omitting the outermost brackets of the antichains. **b)** Information diagrams showing all possible redundancy terms and their nested structure.

5.3 The construction of synergy based partial information decompositions

5.3.1 Proper Synergy

Let us now apply the mereological ideas presented in the previous section to construct a synergy-based PID. To do so, the first question we have to ask is: can we in general express synergistic information I_{syn} as being made up out of certain information atoms $\Pi(f)$? Let us try to work out an answer. Intuitively, the synergy among collections $\mathbf{a}_1, \ldots, \mathbf{a}_m$ should certainly only contain information that is not contained in any individual collection $\mathbf{a}_i$. Otherwise, it would not make sense to call it synergistic. Translating this idea into a constraint on parthood distributions we can say that the synergy should only contain atoms $\Pi(f)$ such that $f(\mathbf{a}_i) = 0$ for any i. Furthermore, it also seems reasonable that the synergy should not contain information that is accessible via some proper subset of sources contained in the $\mathbf{a}_i$. For instance, the synergistic information of sources S_1, S_2, and S_3 about the target should not already be contained in the combination of S_1 and S_2. Also, the synergy between S_1 and the combination (S_2, S_3) should not be accessible if we only

know S_1 and S_2. In terms of parthood distributions we can say the synergy should only contain information atoms $\Pi(f)$ such that $f(\mathbf{b}) = 0$ for all $\mathbf{b} \subset \bigcup \mathbf{a}_i$. This also includes the condition on individual $\mathbf{a}_i$ as a special case. We have now arrived at a negative constraint telling us which atoms are *not* part of the synergy. So the only remaining question is which atoms *are* part of it. Here it appears plausible to demand that if we had access to *all* the collections $\mathbf{a}_i$, then we should obtain the synergistic information they carry about the target. As a parthood constraint this can be expressed as $f(\bigcup \mathbf{a}_i) = 1$. Putting the negative and the positive constraint together this leads to the following relation between synergy I_{syn} and information atoms Π:

$$I_{\text{syn}}(\mathbf{a}_1, \ldots, \mathbf{a}_m : T) = \sum_{\substack{\forall \mathbf{b} \subset \bigcup \mathbf{a}_i f(\mathbf{b}) = 0 \\ f(\bigcup \mathbf{a}_i) = 1}} \Pi(f) \tag{5.14}$$

Now the crucial question is: can this relation be inverted to obtain a solution for all $\Pi(f)$ once a measure of synergy I_{syn} is provided? Unfortunately, the answer is no. The problem is that some of the equations coincide and hence the system is underdetermined. In fact, in the case of three sources, Equation 5.14 only provides four constraints in addition to the consistency equation (11 would be needed). To see this, note first that given the relation above, I_{syn} has to be symmetric, idempotent, and invariant under *subset* removal/addition. Hence, its domain can be reduced to the set of antichains. But there is a further constraint: whenever the union over two antichains is equal, the associated synergy must be equal. Formally,

$$\bigcup \mathbf{a}_i = \bigcup \mathbf{b}_j \rightarrow I_{\text{syn}}(\mathbf{a}_1, \ldots, \mathbf{a}_m : T) = I_{\text{syn}}(\mathbf{b}_1, \ldots, \mathbf{b}_m : T) \textbf{ (Union Condition)} \tag{5.15}$$

Accordingly, there can only be as many independent synergies as there are different non-empty unions (the synergy of the empty set has to be zero). Thus, we are left with seven synergy terms for $n = 3$. Three terms correspond to the singletons $\{i\}$. For these, the condition in Equation 5.14 reduces to $f(\mathbf{a}) = 1$ so that $I_{\text{syn}}(\{i\} : T) = I(S_i : T)$. But this does not provide any constraint beyond the consistency equation. Three further terms correspond to the pairs of sources. And the final term corresponds to the full set of all three sources. It is only the last four terms that genuinely provide novel constraints on the information atoms. They are shown as mereological diagrams in Figure 5.2.

In total, after defining a measure of synergy I_{syn} and given that we also have the consistency equation at our disposal, we are still short seven equations for $n = 3$. An inversion of Equation 5.14 is therefore not possible. Does this mean that there can be no such thing as a synergy-based PID? Not necessarily. It remains a possibility that there are alternative notions of synergistic information, notions that might still capture some, but necessarily not all, of the intuitive properties described above,

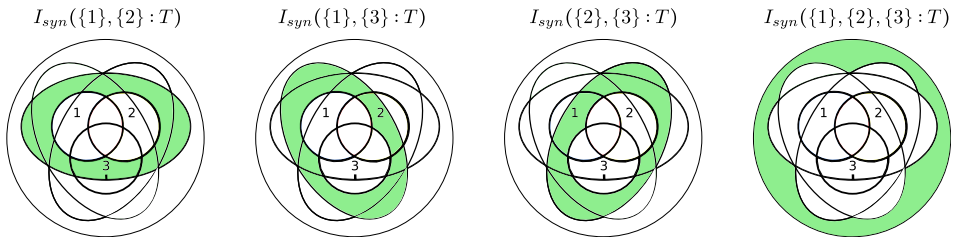

Fig. 5.2: Mereological diagrams of the four independent synergy terms in the $n = 3$ case.

and which allow for the required inversion. We will explore a minimal notion of synergy in the following section.

5.3.2 Weak synergy

Let us strip our concept of synergistic information from anything but it's most essential property: the synergistic information carried by multiple collections of sources about the target should not be accessible via an individual collection $\mathbf{a}_i$. We will call the entirety of the information satisfying this condition the *weak synergy* I_{ws} that collections $\mathbf{a}_1, \ldots, \mathbf{a}_m$ carry about the target [132]. Given this intended meaning of weak synergy it should stand in the following relation to the information atoms:

$$I_{\text{ws}}(\mathbf{a}_1, \ldots, \mathbf{a}_m : T) = \sum_{\forall i f(\mathbf{a}_i)=0} \Pi(f) \tag{5.16}$$

In other words, we sum all atoms that are not part of any individual $I(\mathbf{a}_i : T)$ contribution. Again, in order to determine whether this relation is invertible, we first ask which constraints on I_{ws} are implied by this condition. We obtain the following:

1. $I_{\text{ws}}(\mathbf{a}_1, \ldots, \mathbf{a}_m : T) = I_{\text{ws}}(\mathbf{a}_{\sigma(1)}, \ldots, \mathbf{a}_{\sigma(m)} : T)$ for any permutation σ **(symmetry)** (5.17)
2. If $\mathbf{a}_i \subseteq \mathbf{a}_j$ for $i \neq j$, then $I_{\text{ws}}(\mathbf{a}_1, \ldots, \mathbf{a}_m : T) = I_{\text{ws}}(\mathbf{a}_1, \ldots, \mathbf{a}_{i-1}, \mathbf{a}_{i+1}, \ldots, \mathbf{a}_m : T)$ **(subset invariance)** (5.18)
3. $I_{\text{ws}}(\mathbf{a} : T) = I(\mathbf{a}^C : T|\mathbf{a})$ **(self-synergy)** (5.19)

where $\mathbf{a}^C$ refers to the complement of $\mathbf{a}$. The first two conditions allow us to restrict the weak synergy to the set of antichains. Although this time we can exclude the antichains $\{\}$ and $\{\{1, \ldots, n\}\}$. The reason why the full set does not have to be included is that there is no information atom which is not contained in the information provided by the full set of sources. Accordingly, its weak synergy must be zero. Instead, the set containing the empty set $\{\{\}\}$ has to be included. The information not available if we do not know any source is of course all of the

information in the sources. We refer to the domain of I_{ws} by $\mathcal{S}$. The system of self-synergy equations ensures that the resulting PID satisfies the consistency equation. This is because due to the chain rule for mutual information the conditions

$$I(\mathbf{a} : T) = \sum_{f(\mathbf{a})=1} \Pi(f) \text{ and } I(\mathbf{a}^{\mathbf{C}} : T|\mathbf{a}) = \sum_{f(\mathbf{a})=0} \Pi(f) \tag{5.20}$$

are equivalent.

The relation between weak synergy and information atoms can be rewritten in terms of the ordering on parthood distributions. It is convenient to first turn this lattice upside-down so that the more easily accessible atoms are at the top, i.e. we are considering $(\mathcal{B}, \sqsupseteq)$. By construction, the weak synergy of an antichain $\alpha = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ is equal to all atoms that such that $f(\mathbf{a}_i) = 0$ for all $i = 1, \ldots, m$. But these atoms are precisely the atoms associated with parthood distributions *below and including* the parthood distribution $\tilde{f}_\alpha$ that assigns the value zero to *exactly* all of the $\mathbf{a}_i$ and their subsets (in the upside-down parthood lattice):

$$I_{\text{ws}}(\alpha : T) = \sum_{g \sqsupseteq \tilde{f}_\alpha} \Pi(g) \tag{5.21}$$

All atoms further down in the ordering necessarily also assign the value zero to all $\mathbf{a}_i$ and additionally to some other collections as well (i.e. they are even harder to access). This computation is illustrated in Figure 5.3. What we can see from these considerations is that, just like the redundancies, the weak synergies are nested. The mapping $\alpha \to \tilde{f}_\alpha$ induces a lattice $(\mathcal{S}, \preceq')$ of antichains that describes this nesting. The ordering is given by

$$\alpha \preceq' \beta \Leftrightarrow \tilde{f}_\alpha \sqsupseteq \tilde{g}_\beta \tag{5.22}$$

We will refer to this lattice as the *synergy lattice*. Weak synergies further down in this ordering are contained in synergies higher up. Just like the redundancy ordering, the synergy ordering on antichains also first appeared in a purely order-theoretic work [114] (written in a different but equivalent form). In the context of synergy-based PID it has been utilized by [119] (as “extended constraint lattice”), by [135] (as “information loss lattice”), and most recently by [130] (as "pooling-based lattice"). See Section 5.6 for a discussion of the relation between these approaches and the mereological approach presented here.

By construction the mapping $\alpha \to \tilde{f}_\alpha$ is an isomorphism between the parthood lattice $(\mathcal{B}, \sqsubseteq)$ and the synergy lattice $(\mathcal{S}, \preceq')$. The inverse is given by

$$f \to \tilde{\alpha}_f = \{\mathbf{a} | f(\mathbf{a} = 0) \;\&\; \neg\exists \mathbf{b} \subset \mathbf{a} f(\mathbf{b}) = 0\} \tag{5.23}$$

In other words, $\tilde{\alpha}_f$ consists of the maximal sets $\mathbf{a}$ such that $f(\mathbf{a}) = 0$. Using this isomorphism and the conventions $I_{\text{ws}}(f : T) := I_{\text{ws}}(\tilde{\alpha}_f)$ and $\Pi(\alpha) := \Pi(\tilde{f}_\alpha)$ one can rewrite the relation between weak synergies and atoms as Moebius-Inversions over the parthood and synergy lattices respectively:

$$I_{\text{ws}}(f : T) = \sum_{g \sqsupseteq f} \Pi(g) \qquad\qquad I_{\text{ws}}(\alpha : T) = \sum_{\alpha \preceq' \beta} \Pi(\beta) \tag{5.24}$$

These relations can be inverted once a measure of weak-synergy is specified. We can see here that the construction of weak-synergy-based PIDs proceeds along the same lines as redundancy-based PID. The only difference is that the nesting of weak synergies is described by a different lattice structure. It is important to note that the intended interpretation of the information atoms $\Pi(f)$ remains exactly the same no matter if the PID is induced by a redundancy measure or a weak synergy measure. They still quantify the information that stands in the parthood relations described by f.

Before we proceed to discuss how redundancy and weak synergy are special cases of a more general construction of PID base-concepts, we would like to consider an important interpretative point. Note that the formula on the right in (5.24) uses a different way to associate information atoms with antichains that the one used conventionally in the PID literature. In the standard way each information atom is associated with an antichain α in the redundancy lattice via the isomorphism $f \to \alpha_f$ we considered in Section 5.2. Given such an antichain $\alpha = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ the associated information atom $\Pi(\alpha)$ is the one which is part of the mutual information provided by any $\mathbf{a}_i$ and any superset thereof while it is not part of the mutual information provided by any other collection. In other words, the antichain tells us what the information atom is part of – leaving it implicit what it is not part of. For instance, the atom $\Pi(\{1\})$ is the information uniquely contained in the first source. But there is also an alternative way that uses the synergy related isomorphism $f \to \tilde{\alpha}_f$, associating each atom with an antichain in the synergy lattice. Here the antichain tells us what the corresponding information atom is *not* part of – leaving it implicit what it is part of. In this interpretation the information atom $\tilde{\Pi}(\alpha)$ is not part of the mutual information provided by any $\mathbf{a}_i$ and any subset thereof while it is part of the information provided by any other collection. Accordingly, the unique information of the first source is $\tilde{\Pi}(\{1, \ldots, n\} \backslash \{1\})$ in this notation. It is of course straightforward to convert the two notations by composing the two mappings:

$$\tilde{\Pi}(\alpha) = \Pi(\beta_{\tilde{f}_\alpha}) \qquad\qquad \Pi(\alpha) = \tilde{\Pi}(\tilde{\beta}_{f_\alpha}) \tag{5.25}$$

It is merely a matter of convenience which notation is used. However, when it comes to the interpretation of the information atoms it is important to be clear on this point.

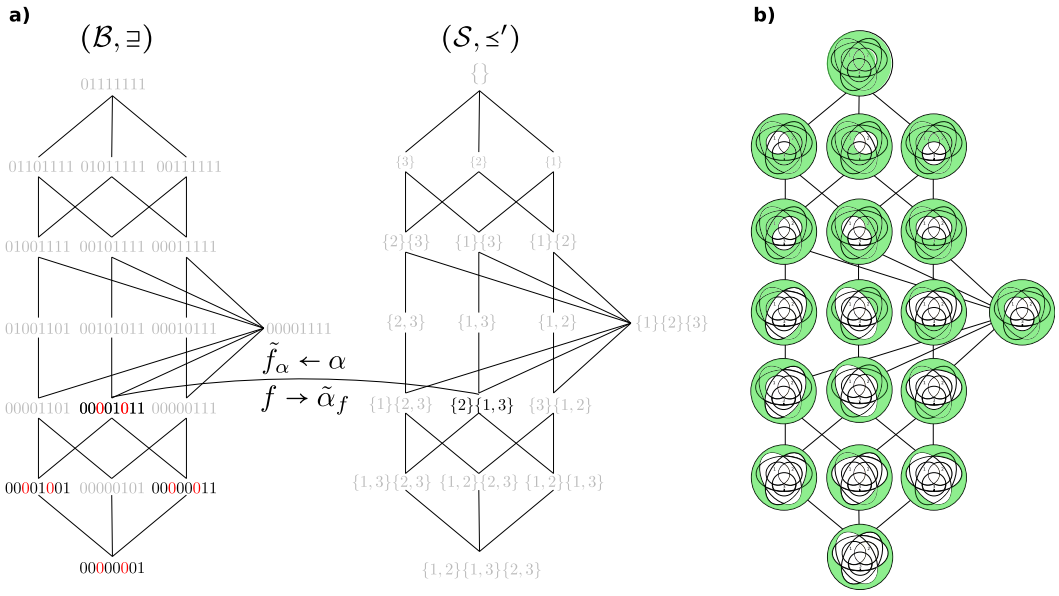

Fig. 5.3: **a)** Parthood and synergy lattices for $n = 3$ sources. There is an isomorphism between the lattices such that the weak synergy associated with a node in the synergy lattice is equal to the sum of atoms associated with parthood distributions above and including the corresponding node in the parthood lattice. This is shown for the antichain $\{2\}\{1,3\}$. Note that we adhere to the standard convention of omitting the outermost brackets of the antichains. **b)** Mereological information diagrams depicting the different synergy terms.

5.4 The logical organization of PID base-concepts

The construction of weak synergy and redundancy suggests a more general scheme for defining composite information measures. This construction defines the information associated with an antichain α in terms of sufficient, necessary, insufficient or unnecessary conditions on parthood or non-parthood with respect to either subsets or supersets of the $\mathbf{a} \in \alpha$. In the case of weak synergy, we are asking for all information such that it is a *sufficient condition* for an atom to be included in this information that *it is not part* of the information provided by any *subset* of the $\mathbf{a} \in \alpha$. We can rewrite the parthood condition of weak synergy (i.e. the condition f has to satisfy so that $\Pi(f)$ is included in the weak synergy associated with α) to make this more explicit. Setting $[n] = \{1, \ldots, n\}$:

$$\forall\, \mathbf{b} \subseteq [n] : \exists\, \mathbf{a} \in \alpha\ \mathbf{b} \subseteq \mathbf{a} \rightarrow f(\mathbf{b}) = 0 \tag{5.26}$$

which picks out exactly the same information atoms for each α as the condition $\forall\, \mathbf{a} \in \alpha : f(\mathbf{a}) = 0$. Similarly, in the case of redundant information we are asking for all information such that it is a *sufficient condition* for an atom to be included in this information that *it is part* of the information provided by any *superset* of the $\mathbf{a} \in \alpha$:

$$\forall\, \mathbf{b} \subseteq [n] : \exists\, \mathbf{a} \in \alpha\ \mathbf{b} \supseteq \mathbf{a} \to f(\mathbf{b}) = 1 \tag{5.27}$$

which picks out the same atoms as $\forall\, \mathbf{a} \in \alpha : f(\mathbf{a}) = 1$. In total the logical construction allows 16 possibilities. Before studying them in detail we would like to introduce a notion which will turn out to be very useful in the subsequent analysis.

Definition 4 (Partner measure)**.** *Let $\mathcal{A}^*, \mathcal{A}^{**} \subseteq \mathbb{A}$. Two information measures $I^* : \mathcal{A}^* \to \mathbb{R}$ and $I^{**} : \mathcal{A}^{**} \to \mathbb{R}$ are partner measures just in case there is a bijective mapping $\varphi : \mathcal{A}^* \to \mathcal{A}^{**}$ such that $I^*(\alpha : T) = I^{**}(\varphi(\alpha) : T) \forall \alpha \in \mathcal{A}^*$.*

where $\mathbb{A}$ is the set of *all* antichains of the partial order $([n], \subseteq)$, i.e. including both $\{\{\}\}$ and $\{\{1, \ldots, n\}\}$ as well as the empty set $\{\}$.

Partner measures quantify the same kind of information but viewed from the perspective of different collections. An example would be weak synergy and the "restricted information" we introduced in [132]. The information we cannot get from any individual $\mathbf{a} \in \alpha$ (weak synergy) is exactly the information we can *only* get from other collections (i.e. the information restricted to these other collections), where the “other” collections are all non-subsets of the $\mathbf{a} \in \alpha$. This is illustrated in the top right corner of Figure 5.4. In the following, we will be interested specifically in partner measures with respect to the following two mappings between antichains

$$\alpha \longmapsto \bar{\alpha} = \min_{\subseteq} \left(\{ \mathbf{b} \in [n] \mid \neg \exists\, \mathbf{a} \in \alpha : \mathbf{b} \subseteq \mathbf{a} \} \right) \tag{5.28}$$

$$\alpha \longmapsto \underline{\alpha} = \max_{\subseteq} \left(\{ \mathbf{b} \in [n] \mid \neg \exists\, \mathbf{a} \in \alpha : \mathbf{b} \supseteq \mathbf{a} \} \right) \tag{5.29}$$

The first mapping collects the minimal non-subsets of the collections in α and the second one collects all the maximal non-supersets of these collections. Restricted information is a partner measure of weak synergy with respect to the first of the two mappings. Because the two mappings are inverses of each other (for proof see 5.8.1), weak synergy is a partner measure of restricted information with respect to the second mapping. Figure 5.5 shows the two mappings for $n = 2$.

Let us now consider all the possible cases of the general construction of information measures described above:

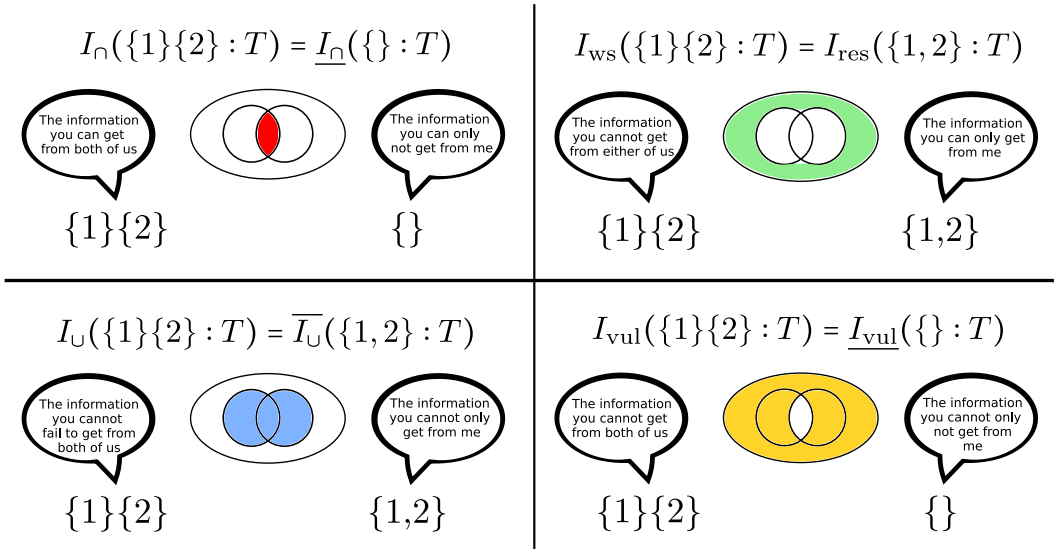

Fig. 5.4: Intuitive interpretation of partner measures in the case $n = 2$. *Top left:* redundant information and its partner measure. The information which is redundant to both sources, $I_\cap(\{1\}\{2\} : T)$, is the information that we can only not get if we do not know any source, i.e. $\underline{I_\cap}(\{\} : T)$. *Top right:* weak synergy and its partner measure. The information we cannot get from either source individually, $I_{ws}(\{1\}\{2\} : T)$, is the information we can only get if we know both sources at the same time, i.e. the information restricted to the full set of sources $I_{res}(\{1,2\} : T) = \overline{I_{ws}}(\{1,2\} : T)$. *Bottom left:* union information and its partner measure. The union information, $I_\cup(\{1\},\{2\} : T)$, is the information we cannot fail to get from both individual sources. Or in other words, it is all information we can get from at least one individual source. This can equivalently be described as the information, $\overline{I_\cup}(\{1,2\} : T)$, we cannot *only* get if we know both sources, i.e. for each component of the union information there is a way to access it that does not require full knowledge of both sources. *Bottom right:* vulnerable information and its partner measure. The vulnerable information, $I_{vul}(\{1\}\{2\} : T)$, is all information we cannot get from both sources. This means that for each component of the vulnerable information there is a scenario in which we fail to obtain it *other than the scenario in which we do not know any of the sources*. Therefore, it is the information we cannot only not get from the empty set of sources, i.e. $\underline{I_{vul}}(\{\} : T)$.

Sufficient Conditions There are four conditions saying that being a subset/superset of some collection $\mathbf{a} \in \alpha$ is sufficient for parthood/non-parthood:

$$\forall \mathbf{b} \subseteq [n] : \exists \mathbf{a} \in \alpha \ \mathbf{b} \supseteq \mathbf{a} \to f(\mathbf{b}) = 1 \quad \forall \mathbf{b} \subseteq [n] : \exists \mathbf{a} \in \alpha \ \mathbf{b} \subseteq \mathbf{a} \to f(\mathbf{b}) = 0 \tag{5.30}$$

$$\forall \mathbf{b} \subseteq [n] : \exists \mathbf{a} \in \alpha \ \mathbf{b} \subseteq \mathbf{a} \to f(\mathbf{b}) = 1 \quad \forall \mathbf{b} \subseteq [n] : \exists \mathbf{a} \in \alpha \ \mathbf{b} \supseteq \mathbf{a} \to f(\mathbf{b}) = 0 \tag{5.31}$$

We already discussed the first two conditions above. They correspond to redundancy and weak synergy respectively. The second two conditions are trivial. The first one because all parthood distributions satisfy $f(\{\}) = 0$. Thus, there is always a $\mathbf{b}$ for which the antecedent is true while the consequent is false. Accordingly, no information is included in the information described by the condition. Phrased

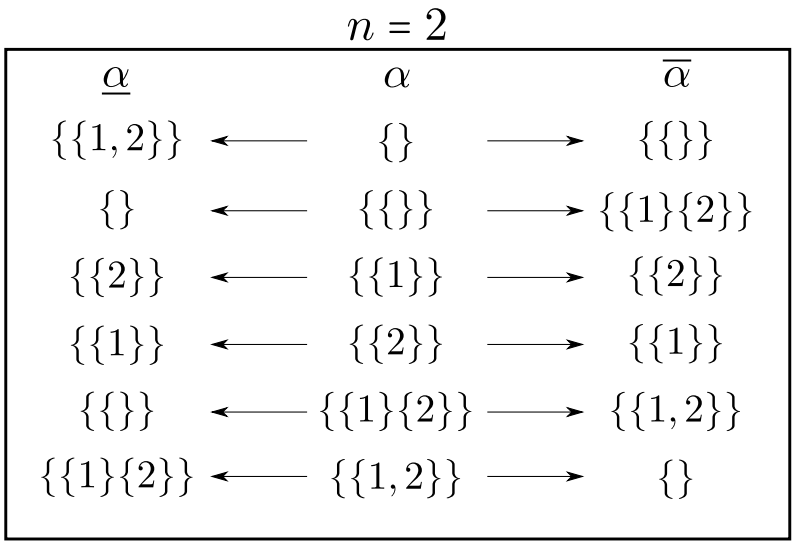

Fig. 5.5: Mappings 5.28 and 5.29 for $n = 2$. Antichains $\alpha \in \mathcal{A}$ (middle column) are mapped to either $\underline{\alpha} \in \mathcal{A}$ (left column) or $\overline{\alpha} \in \mathcal{A}$ (right column).

differently, it is never sufficient for an information atom to be part of $I(\mathbf{b} : T)$ that $\mathbf{b}$ is a subset of some $\mathbf{a} \in \alpha$. Analogously, the second condition does not include any information atom because we always have $f(\{1, \ldots, n\}) = 1$. It is never sufficient for an information not to be part of $I(\mathbf{b} : T)$ that $\mathbf{b}$ is a superset of some $\mathbf{a} \in \alpha$.

Necessary Conditions The following conditions express that being a subset/superset of some $\mathbf{a} \in \alpha$ is necessary for parthood/non-parthood:

$$\forall \mathbf{b} \subseteq [n] : \neg\exists \mathbf{a} \in \alpha \; \mathbf{b} \supseteq \mathbf{a} \rightarrow f(\mathbf{b}) = 1 \qquad \forall \mathbf{b} \subseteq [n] : \neg\exists \mathbf{a} \in \alpha \; \mathbf{b} \subseteq \mathbf{a} \rightarrow f(\mathbf{b}) = 1 \tag{5.32}$$

$$\forall \mathbf{b} \subseteq [n] : \neg\exists \; \mathbf{a} \in \alpha \; \mathbf{b} \supseteq \mathbf{a} \rightarrow f(\mathbf{b}) = 0 \qquad \forall \mathbf{b} \subseteq [n] : \neg\exists \mathbf{a} \in \alpha \; \mathbf{b} \subseteq \mathbf{a} \rightarrow f(\mathbf{b}) = 0 \tag{5.33}$$

The first condition in 5.32 says that being a superset of some $\mathbf{a} \in \alpha$ is necessary for *non-parthood*, i.e. in order for an atom to not be part of $I(\mathbf{b} : T)$ it must be the case that $\mathbf{b}$ is a superset of some $\mathbf{a} \in \alpha$. But this is never true. There is always a non-superset satisfying $f(\mathbf{b}) = 0$, namely $\mathbf{b} = \{\}$. Accordingly, the condition picks out none of the information atoms with the sole exception of the antichain $\alpha = \{\{\}\}$. Here it trivially picks out all atoms.

The second condition in 5.32 is the partner measure of redundancy $\underline{I_\cap}$ because from the perspective of the $\underline{\mathbf{a}} \in \underline{\alpha}$ the non-subsets are exactly the $\mathbf{a} \in \alpha$ and their supersets so that $\underline{I_\cap}(\underline{\alpha} : T) = I_\cap(\alpha : T)$. It says that being a subset of some $\mathbf{a} \in \alpha$ is necessary for non-parthood. Accordingly, the information picked out by the condition must be redundant with respect to all non-subsets of the collections at which $\underline{I_\cap}$ is evaluated. For an illustration see the top left corner of Figure 5.4.

In 5.33, the first condition describes the partner measure of weak synergy $\overline{I_{\text{ws}}}$ because from the perspective of the $\overline{\mathbf{a}} \in \overline{\alpha}$ the non-supersets are exactly the $\mathbf{a} \in \alpha$ and their subsets so that $\overline{I_{\text{ws}}}(\overline{\alpha} : T) = I_{\text{ws}}(\alpha : T)$. It says that being a superset of some $\mathbf{a} \in \alpha$ is necessary for *parthood*, i.e. it captures information that may only be contained in the collections at which it is evaluated and their supersets. This is the *restricted information* discussed above (see also top right corner of Figure 5.4).

The second condition in 5.33 says that being a subset of some $\mathbf{a} \in \alpha$ is necessary for *parthood*. But this is never the case. There is always a non-subset satisfying $f(\mathbf{b}) = 1$, namely $\mathbf{b} = \{1, \dots, n\}$. Accordingly, the condition picks out none of the information atoms with the sole exception of the antichain $\alpha = \{\{1, \dots, n\}\}$ where it trivially picks out all atoms.

Insufficient Conditions The conditions expressing that being a subset/superset of some $\mathbf{a} \in \alpha$ is insufficient for parthood/non-parthood are

$$\neg(\forall \mathbf{b} \subseteq [n] : \exists \mathbf{a} \in \alpha \; \mathbf{b} \supseteq \mathbf{a} \to f(\mathbf{b}) = 1) \quad \neg(\forall \mathbf{b} \subseteq [n] : \exists \mathbf{a} \in \alpha \; \mathbf{b} \subseteq \mathbf{a} \to f(\mathbf{b}) = 1) \tag{5.34}$$

$$\neg(\forall \mathbf{b} \subseteq [n] : \exists \mathbf{a} \in \alpha \; \mathbf{b} \supseteq \mathbf{a} \to f(\mathbf{b}) = 0) \quad \neg(\forall \mathbf{b} \subseteq [n] : \exists \mathbf{a} \in \alpha \; \mathbf{b} \subseteq \mathbf{a} \to f(\mathbf{b}) = 0) \tag{5.35}$$

The first condition in (5.34) leads to a measure of information that has not been described in the literature before. Intuitively, it describes the "the information we do not get from at least one $\mathbf{a} \in \alpha$". One might call this *vulnerable information* because it is not completely redundant with respect to the $\mathbf{a} \in \alpha$ and hence may be lost if we loose access to some of these collections (or is not contained in any of them in the first place). It is the complement of the redundancy. The second condition in (5.34) is trivial. It includes all atoms because there is always a subset of the $\mathbf{b} \in \alpha$ for which $f(\mathbf{b}) = 0$, namely $\mathbf{b} = \{\}$. Similarly, the first condition in (5.35) includes all atoms because there is always a superset of the $\mathbf{b} \in \alpha$ for which $f(\mathbf{b}) = 1$, namely $\mathbf{b} = \{1, \dots, n\}$. The second condition in (5.35) describes the union information, i.e. the information we can obtain from at least one $\mathbf{a} \in \alpha$.

Unnecessary Conditions Finally, there are four conditions saying that being a subset/superset of some $\mathbf{a} \in \alpha$ is unnecessary for parthood/non-parthood:

$$\neg(\forall \mathbf{b} \subseteq [n] : \neg\exists \mathbf{a} \in \alpha \; \mathbf{b} \supseteq \mathbf{a} \to f(\mathbf{b}) = 1) \quad \neg(\forall \mathbf{b} \subseteq [n] : \neg\exists \mathbf{a} \in \alpha \; \mathbf{b} \subseteq \mathbf{a} \to f(\mathbf{b}) = 1) \tag{5.36}$$

$$\neg(\forall \mathbf{b} \subseteq [n] : \neg\exists \; \mathbf{a} \in \alpha \; \mathbf{b} \supseteq \mathbf{a} \to f(\mathbf{b}) = 0) \quad \neg(\forall \mathbf{b} \subseteq [n] : \neg\exists \mathbf{a} \in \alpha \; \mathbf{b} \subseteq \mathbf{a} \to f(\mathbf{b}) = 0) \tag{5.37}$$

The first condition in 5.36 is trivial. It says that being a superset of some $\mathbf{a} \in \alpha$ is not necessary for non-parthood. But this is true for all antichains and information atoms because there always a non-superset for which $(f(\mathbf{b})) = 0$, namely $\mathbf{b} = \{\}$. The only

exception is $\alpha = \{\{\}\}$ for which the condition trivially picks out no information atom. The second condition in 5.36 is the partner measure $\underline{I_{\text{vul}}}$ of vulnerable information because from the perspective of $\underline{\alpha}$ the non-subsets are exactly the $\mathbf{a} \in \alpha$ and their supersets so that $\underline{I_{\text{vul}}}(\underline{\alpha} : T) = I_{\text{vul}}(\alpha : T)$. For an intuitive description of vulnerable information and its partner measure see the bottom right corner of Figure 5.4.

The first condition in 5.37 is the partner measure $\overline{I_\cup}$ of union information because from the perspective of $\overline{\alpha}$ the non-supersets are exactly the $\mathbf{a} \in \alpha$ and their subsets so that $\overline{I_\cup}(\overline{\alpha} : T) = I_\cup(\alpha : T)$. For an intuitive description of union information and its partner measure see the bottom left corner of Figure 5.4. The second condition in 5.37 is trivial because it says that being a subset of some $\mathbf{a} \in \alpha$ is not necessary for parthood. But this is true for all antichains and information atoms since there is always a non-subset for which $f(\mathbf{b}) = 1$, namely $\mathbf{b} = \{1, \ldots, n\}$. The only exception is $\alpha = \{\{1, \ldots, n\}\}$ where the condition trivially picks out no atom.

So in total we obtain four pairs of partner measures as shown in Figure 5.6 for the case $n = 2$. The Figure also locates previous PID approaches within this scheme. Thus far, there has been no proposal utilizing vulnerable information as a PID base-concept. Furthermore, all proposals in the literature are based on $I_\cap$, $I_\cup$, or I_{ws} rather than their partner measures. Some comments are in order in particular about the weak synergy quadrant: The measure of "synergistic disclosure" by Rosas et al [119] is very close in spirit to what we have called weak synergy here but only leads to a standard PID when it is modified appropriately. This is discussed in Section 5.6.1 below. The approach by Perrone & Ay [82] does not attempt to construct a PID but rather a decomposition of joint mutual information into interactions of orders 1 to n. In the two-sources case this amounts to defining union information and the synergy atom which is why we included it in parentheses.

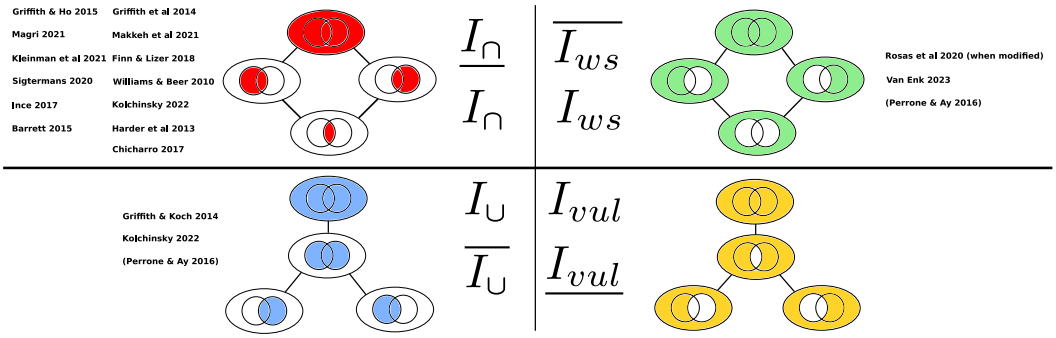

Fig. 5.6: Scheme of four equivalence classes of partner measures. Previous PID approaches are categorized in the appropriate quadrants.

Before discussing the implied properties and associated lattices of the different base-concepts we would like to briefly address a base-concept that we have not considered so far: unique information. This has only been utilized in the two-sources case [83, 128, 129]. However, we argued in [132] that one may generalize the concept so that it becomes a base-concept in the general case as well. Given collections of source variables α one may think of the unique information associated with these collections as "the information contained in all of the $\mathbf{a} \in \alpha$ but nowhere else". In other words, it consists of the information atoms $\Pi(f)$ where $f(\mathbf{b}) = 1$ if $\mathbf{b} \supseteq \mathbf{a}$ for some $\mathbf{a} \in \alpha$ and $f(\mathbf{b}) = 0$ otherwise. There is only one such information atom, namely the atom $\Pi(f_\alpha)$ so that we have $I_{\text{unq}}(\alpha : T) = \Pi(f_\alpha) = \Pi(\alpha)$ (see Section 5.3.2 above for an explanation of this notation). Hence, a unique information based PID amounts to defining the information atoms directly. Unique information can also be described by a logical condition similar to the ones we discussed above. It is captured by a *sufficient and necessary* condition with respect to parthood in supersets of the $\mathbf{a}_i$:

$$\forall \mathbf{b} \subseteq [n] : \exists \mathbf{a} \in \alpha \mathbf{b} \supseteq \mathbf{a} \rightarrow f(\mathbf{b}) = 1 \;\&\; \forall \mathbf{b} \subseteq [n] : \neg\exists\, \mathbf{a} \in \alpha\; \mathbf{b} \supseteq \mathbf{a} \rightarrow f(\mathbf{b}) = 0 \tag{5.38}$$

This condition is the logical conjunction of the conditions for $I_\cap$ and $\overline{I_{\text{ws}}}$. It also has a natural partner measure arising from the conjunction of the I_{ws} and $\underline{I_\cap}$ conditions which amounts to a *sufficient and necessary* condition with respect to non-parthood in subsets of the $\mathbf{a}_i$:

$$\forall \mathbf{b} \subseteq [n] : \exists \mathbf{a} \in \alpha\; \mathbf{b} \subseteq \mathbf{a} \rightarrow f(\mathbf{b}) = 0 \;\&\; \forall \mathbf{b} \subseteq [n] : \neg\exists \mathbf{a} \in \alpha\; \mathbf{b} \subseteq \mathbf{a} \rightarrow f(\mathbf{b}) = 1 \tag{5.39}$$

This describes the partner measure $\underline{I_{\text{unq}}}$ and can be interpreted as "the information we do not get from any of the $\mathbf{a} \in \alpha$ but anywhere else". It satisfies $\underline{I_{\text{unq}}}(\alpha : T) = \Pi(\tilde{f}_\alpha) = \tilde{\Pi}(\alpha)$ (see Section 5.3.2 above for an explanation of this notation).

5.5 Properties and Lattices

Each of the information measures discussed in the previous section is associated with a particular lattice (or semi-lattice) describing its nested structure (except of course unique information since it is not nested and simply has to satisfy the consistency equation 5.4). For redundancy and weak synergy these are the lattices $(\mathcal{A}, \preceq)$ and $(\mathcal{S}, \preceq')$ as introduced in Section 5.2 and 5.3. Furthermore, each information measure has a range of fundamental properties following from their characteristic parthood conditions. For redundancy and weak synergy these are the above Equations 5.6-5.7 and 5.17-5.18, respectively. The corresponding lattices and properties of the other base-concepts can be derived easily utilizing their relations to redundancy and weak synergy as well as the mappings $\alpha \rightarrow \overline{\alpha}$ and $\alpha \rightarrow \underline{\alpha}$.

The redundancy partner $\underline{I_\cap}$: The domain of the $\underline{I_\cap}$ is the image of $\mathcal{A}$ under $\alpha \to \underline{\alpha}$, i.e. $\underline{\mathcal{A}} = \mathbb{A}\backslash\{\{\{1,\ldots,n\}\},\{\}\} = \mathcal{S}$. In order to find the ordering relation note that

$$\underline{I_\cap}(\alpha : T) = \sum_{g \sqsubseteq \tilde{f}_\alpha} \Pi(g) \tag{5.40}$$

The left hand side expresses the information at most not contained in the $\mathbf{a} \in \alpha$ and their subsets. But this is equal to the information atom $\Pi(\tilde{f}_\alpha)$, which is not contained *exactly* in all $\mathbf{a} \in \alpha$ and subsets thereof, plus all information atoms further down the parthood lattice, i.e. all more accessible atoms. Hence, the appropriate ordering relation is the inverted weak synergy ordering (compare Equation 5.21 above) so that the lattice for $\underline{I_\cap}$ is $(\mathcal{S}, \succeq')$. In other words, using $\underline{I_\cap}$ as a base-concept amounts to performing an upwards Moebius-Inversion over the synergy lattice. $\underline{I_\cap}$ is symmetric, subset-invariant and satisfies the condition

$$\underline{I_\cap}([n]\backslash\{i_1\},\ldots,[n]\backslash\{i_m\} : T) = I(\{i_1,\ldots,i_m\} : T) \tag{5.41}$$

The weak synergy partner $\overline{I_{\text{ws}}} = I_{\text{res}}$: Analogously, the domain of $\overline{I_{\text{ws}}}$ is the image of $\mathcal{S}$ under $\alpha \to \overline{\alpha}$, i.e. $\overline{\mathcal{S}} = \mathbb{A}\backslash\{\{\{\}\},\{\}\} = \mathcal{A}$, equipped with the inverted redundancy ordering because

$$\overline{I_{\text{ws}}}(\alpha : T) = \sum_{g \sqsupseteq f_\alpha} \Pi(g) \tag{5.42}$$

The information at most contained in a superset of the $\mathbf{a} \in \alpha$ is equal to the information atom $\Pi(f_\alpha)$ which is contained *exactly* in all $\mathbf{a} \in \alpha$ and their supersets, plus all information atoms further down the parthood lattice, i.e. all even less accessible atoms. Hence, the nesting is described by the lattice $(\mathcal{A}, \succeq)$. In other words, using $\overline{I_{\text{ws}}}$ as a base-concept amounts to performing an upwards Moebius-Inversion over the redundancy lattice. $\overline{I_{\text{ws}}}$ is symmetric, superset-invariant and satisfies the condition

$$\overline{I_{\text{ws}}}(\{i_1\},\ldots,\{i_m\} : T) = I(\{i_1,\ldots,i_m\} : T | \{i_1,\ldots,i_m\}^C) \tag{5.43}$$

Union information and its partner: Since union information is the complement of weak synergy, i.e. the atoms summed over to obtain the union information are exactly the atoms not summed over to obtain the weak synergy and vice versa, the nesting of union information terms must be described by the inverted weak synergy ordering. There is one union information for every antichain in the synergy lattice except for $\{\{\}\}$ which captures all information if the weak synergy is applied to it and hence captures no information if the union information is applied to it. Instead the

antichain $\{1, \ldots, n\}$ is included because it captures no information with respect to weak synergy and hence all information with respect to union information. Thus the nesting of union information terms is described by the semi-lattice $(\mathcal{A}, \geq')$. It is not a full lattice because it has multiple lowest elements. See Figure 5.7 for the case $n = 3$. The solution for the information atoms is not a Moebius-Inversion. The system of equations is still invertible because it is merely an equivalence transformation of the weak synergy system. Given a specific measure of union information $I^*_\cup$ the solution for the information atoms is equal to their solution in the weak synergy system where we set

$$I^*_{\text{ws}}(\mathbf{a}_1, \ldots, \mathbf{a}_m : T) := I(\{1, \ldots, n\} : T) - I^*_\cup(\mathbf{a}_1, \ldots, \mathbf{a}_m : T) \tag{5.44}$$

Union information is symmetric, subset-invariant, and satisfies

$$I_\cup(\mathbf{a} : T) = I(\mathbf{a} : T) \tag{5.45}$$

The domain of the partner measure of union information $\overline{I_\cup}$ is the image of $\mathcal{A}$ under $\alpha \to \overline{\alpha}$, i.e. $\overline{A} = \mathbb{A}\backslash\{\{\{\}\}, \{\{1\}, \{2\}\}\}$ which is not equal to any domain we have considered before. Since it is the complement of restricted information $I_{\text{res}} = \overline{I_{\text{ws}}}$, its nesting is described by the semi-lattice $(\overline{\mathcal{A}}, \preceq)$. $\overline{I_\cup}$ is symmetric, superset-invariant and satisfies

$$\overline{I_\cup}(\{i_1\}, \ldots, \{i_m\} : T) = I(\{i_1, \ldots, i_m\} : T) \tag{5.46}$$

Vulnerable information and its partner: Since vulnerable information is the complement of redundancy the nesting of vulnerable information terms must be described by the inverted redundancy ordering. Anologously to the the discussion of union information we conclude that the domain of vulnerable information is $\mathbb{A}\backslash\{\{\}, \{1, \ldots, n\}\} = \mathcal{S}$. Hence, the nesting of vulnerable information terms is described by the semi-lattice $(\mathcal{S}, \geq)$. Again, the solution for the information atoms does not have the structure of a Moebius-Inversion. See Figure 5.8 for the case $n = 3$. The underlying system of equations is an equivalence transformation of the redundancy system and is therefore solvable. Given a specific measure of vulnerable information I^*_{vul} the solution for the information atoms is equal to their solution in the redundancy system where we set

$$I^*_\cap(\mathbf{a}_1, \ldots, \mathbf{a}_m : T) := I(\{1, \ldots, n\} : T) - I^*_{\text{vul}}(\mathbf{a}_1, \ldots, \mathbf{a}_m : T) \tag{5.47}$$

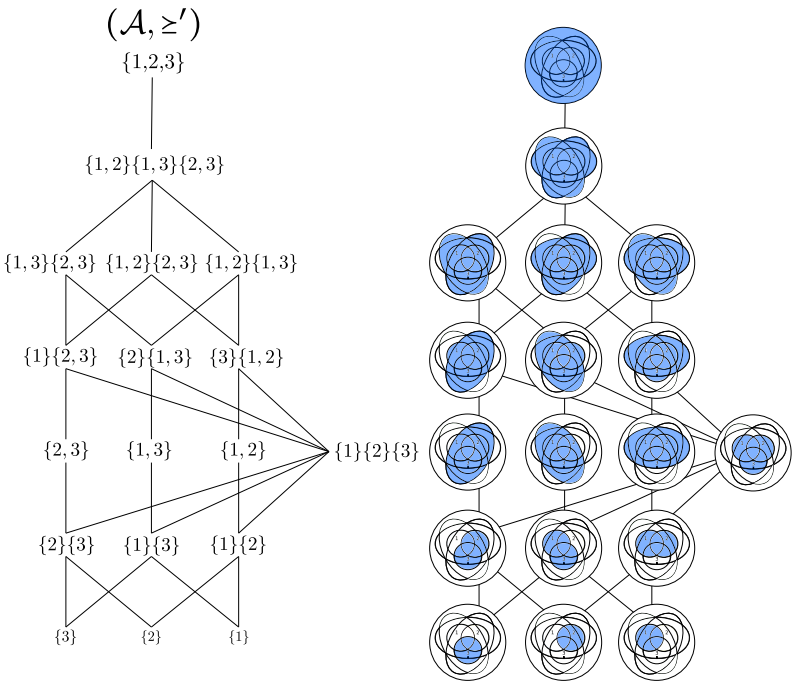

Fig. 5.7: Left: Union information semi-lattice for $n = 3$ sources. Right: Mereological information diagrams depicting the different union information terms.

Vulnerable information is symmetric, superset-invariant, and satisfies

$$I_{\text{vul}}(\mathbf{a} : T) = I(\mathbf{a}^C : T|\mathbf{a}) \tag{5.48}$$

The partner of vulnerable information $\underline{I_{\text{vul}}}$ is defined on the domain

$$\underline{\mathcal{S}} = \mathbb{A}\backslash\{\{\{1\}, \ldots, \{n\}\}, \{\{1, \ldots, n\}\}\} \tag{5.49}$$

which again is different from those we considered before. Since $\underline{I_{\text{vul}}}$ is the complement of $\underline{I_\cap}$ its semi-lattice must be $(\underline{\mathcal{S}}, \preceq')$. $\underline{I_{\text{vul}}}$ is symmetric, subset-invariant and satisfies

$$\underline{I_{\text{vul}}}([n]\backslash\{i_1\}, \ldots, [n]\backslash\{i_m\} : T) = I(\{i_1, \ldots, i_m\}^C : T|\{i_1, \ldots, i_m\}) \tag{5.50}$$

Inclusion-Exclusion The logical conditions defining the different base-concepts do not only entail their individual properties as discussed above. Since each of them stands in an invertible relation to the information atoms, fixing one of them automatically fixes the others as well. We would like to illustrate this for the base-concepts of redundancy and union information. Based on their defining logical

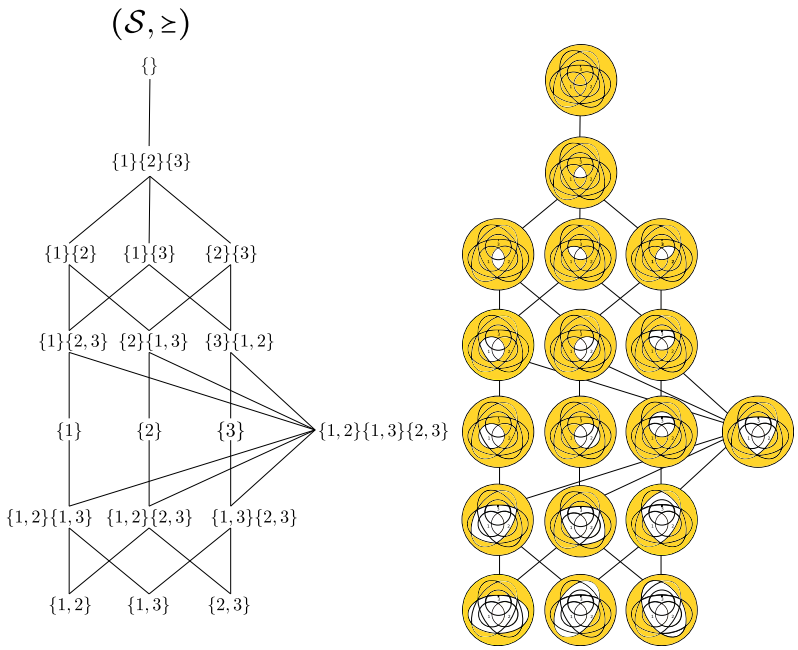

Fig. 5.8: Left: Vulnerable information semi-lattice for $n = 3$ sources. Right: Mereological information diagrams depicting the different vulnerable information terms.

conditions on parthood relations these base-concepts must stand in an inclusion-exclusion relationship:

$$
\begin{aligned}
I_\cup(\mathbf{a}_1, \ldots, \mathbf{a}_m : T) &= \sum_{1 \leqslant i \leqslant m} \sum_{f(\mathbf{a}_i)=1} \Pi(f) - \sum_{1 \leqslant i < j \leqslant m} \sum_{\substack{f(\mathbf{a}_i)=1 \\ f(\mathbf{a}_j)=1}} \Pi(f) + \sum_{1 \leqslant i < j < k \leqslant m} \sum_{\substack{f(\mathbf{a}_i)=1 \\ f(\mathbf{a}_j)=1 \\ f(\mathbf{a}_k)=1}} \Pi(f) - \ldots \\
&= \sum_{1 \leqslant i \leqslant m} I_\cap(\mathbf{a}_i : T) - \sum_{1 \leqslant i < j \leqslant m} I_\cap(\mathbf{a}_i, \mathbf{a}_j : T) + \sum_{1 \leqslant i < j < k \leqslant m} I_\cap(\mathbf{a}_i, \mathbf{a}_j, \mathbf{a}_k : T) - \ldots
\end{aligned}
$$

To see why the first equation is true consider its first summand. It involves all the information atoms that are part of at least one $I(\mathbf{a}_i : T)$. These are by construction exactly the atoms making up the union information $I_\cup(\mathbf{a}_1, \ldots, \mathbf{a}_m : T)$. However, some of these atoms are counted multiple times in the first summand. In particular, if such an atom is part of k mutual information terms $I(\mathbf{a}_i : T)$, it will be counted k times. So the remaining summands must make sure that each atom is only counted exactly once. This is true for the following reason: take any information atom $\Pi(f)$ appearing in the first summand and assume it is part of k mutual information terms $I(\mathbf{a}_i : T)$. It is counted k times by the first summand, $\binom{k}{2}$ times by the second summand, $\binom{k}{3}$ times by the third one, and so on until the k-th summand which counts it one time. So in total it is counted $\sum_{i=1}^{k} (-1)^{i+1} \binom{k}{i} = 1$ times, as desired.

5.6 Relation to previous approaches

5.6.1 Modified Synergistic Disclosure

Rosas et al [119] recently introduced a well motivated measure of synergistic information that is conceptually very similar to the notion of weak synergy introduced in the previous section. The measure is based on the idea of *synergistic observables*. Given an antichain $\alpha = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ an α-synergistic observable V is a univariate random variable such that $I(V : \mathbf{a}_i) = 0$ for $i = 1, \ldots, m$. In other words, a synergistic observable does not contain any information about an individual collection $\mathbf{a}_i$. The synergy of source collections $\mathbf{a}_1, \ldots, \mathbf{a}_m$ is then defined as the supremum of the information provided by synergistic observables that additionally satisfy the Markov condition $V - S - T$:

$$I_{SD}(\alpha : T) = \sup_{\substack{\text{V is } \alpha-synergistic \\ V-S-T}} I(V : T) \tag{5.51}$$

Intuitively, the Markov condition ensures that the information we are considering is actually contained in the sources so that, once we know them, V does not yield any additional information about the target. One may now introduce synergistic disclosure atoms via a Moebius inversion over the synergy lattice (or, as Rosas et al call it, the "extended constraint lattice") [119]. However, the resulting decomposition is not a standard PID because the *consistency condition* (5.4) is not satisfied. This means that the atoms cannot be interpreted in terms of parthood relations with respect to mutual information terms as described in Section 5.2. For example, we do not obtain any atoms interpretable as unique or redundant information in the case of two sources. This is because there are no two atoms in the decomposition that would necessarily add up to $I(S_1 : T)$. But if there were atoms interpretable as the redundancy between the two sources and unique information of source 1 respectively, then these *should always* add up to $I(S_1 : T)$ (The same problem also arises for $I(S_2 : T)$).

In order to construct a standard PID out of the synergistic disclosure measure one may however replace the self-disclosures with the appropriate conditional mutual information terms to enforce the consistency condition (5.4) to be be satisfied. The resulting modified synergistic disclosure measure is defined as:

$$I_{MSD}(\alpha : T) = \begin{cases} I_{SD}(\alpha : T) & \text{if } |\alpha| \geqslant 2 \\ I(T : \mathbf{a}_1^C | \mathbf{a}_1) & \text{if } |\alpha| = 1. \end{cases} \tag{5.52}$$

5.6.2 Loss and Gain Lattices

In a 2017 paper Chicharro and Panzeri [135] introduced partial information decompositions based on what they call information gain and information loss lattices. Structurally, these correspond to what we have here called redundancy lattices and synergy lattices, respectively. And in fact, there is an intuitive way to understand redundancy as an information gain and weak synergy as an information loss: suppose that initially we do not have access to any information source. Now we get access to at least one collection of sources $\mathbf{a}_1, \ldots, \mathbf{a}_m$ (we do not know which). Then the information that we are *guaranteed to gain* should be exactly the redundancy between the $\mathbf{a}_i$. Hence, any redundancy can be described as the guaranteed information gain under such circumstances. Similarly, suppose that initially we have access to all sources. Now we lose access to all except one of the collections $\mathbf{a}_1, \ldots, \mathbf{a}_m$ (we do not know which). Then what we are left with will be exactly the information contained in the remaining collection (which could be any of them) and thus the information we are *guaranteed to lose* should be exactly the information *not* contained in any individual $\mathbf{a}_i$, i.e. their weak synergy. Hence, any weak synergy can be described as the guaranteed information loss under such circumstances.

Indeed, the information loss decomposition is structurally identical to the weak synergy decomposition. It expresses a function of cumulative information loss as a downwards sum over the lattice $(\mathcal{S}, \preceq')$. We would like to point out two differences between the construction of Chicharro and Panzeri and the one presented here: *Firstly*, we start the construction with a characterization of the *components* Π of the mutual information decomposition. Composite information quantities such as redundancy or synergy are introduced via their appropriate relation to these components. The appropriate domains and lattices describing their nested structure can be *derived* from these relations. By contrast the information gain (redundancy-based) and information loss (synergy-based) decompositions are introduced as two separate decompositions involving prima facie distinct sets of information atoms ΔI and ΔL [135]. These are implicitly defined via Moebius-Inversion over the corresponding lattices. This raises the question of how these sets of components are to be interpreted and what their relation should be [135]. Due to the different construction these issues do not arise in the mereological approach. *Secondly*, in the mereological approach redundancy- and synergy-based PID are just special cases of a more general unifying principle allowing the construction of information decompositions in terms of a great variety of base-concepts as discussed in Section 5.4. These base-concepts differ merely in their characteristic logical condition on parthood distributions.

5.7 Conclusion

We presented a general pattern of logical conditions on parthood relations that captures all the PID base-concepts in the literature and that additionally leads to similarly interpretable novel base-concepts. These include in particular the concept of "vulnerable information", i.e. information we cannot obtain from at least one of the source collections at which it is evaluated. This concept may prove useful in a data security context where it the amount of information at risk of being lost since it is not entirely redundant. An interesting fact about vulnerable information is that its nested structure is described only by a semi-lattice and that its underlying system of equations does not have the structure of a Moebius-Inversion. This is how it differs from redundancy or weak synergy. Nonetheless its relationship to the information atoms is invertible and hence leads to a unique PID. The same applies to the concept of union information.

Our construction also leads to "partner measures" for each of the PID base-concepts. These describe the same components of the joint mutual information but from the perspective of different antichains. Accordingly, two partner measures have different domains and (semi-)lattices describing their nested structure. One insight to be gained from this is that a synergy-based PID (in the form of its partner measure $\overline{I_{\text{ws}}}$) is obtainable via an upwards Moebius-Inversion on the redundancy lattice while a redundancy-based PID (in the form of its partner $\underline{I_{\cap}}$) can be obtained via an upwards Moebius-Inversion over the synergy-lattice. Overall, the unifying analysis presented here provides, on the one hand, more theoretical options for inducing PIDs that might be particularly suitable for certain applications contexts and, on the other, it lays the groundwork for detailed theoretical studies into the compatibility between properties of different base-concepts as functions of the underlying joint distribution. This latter point will be a particularly intriguing topic for future studies.

5.8 Appendix

5.8.1 Proof that the partner measure mappings are inverses of each other

First note that the non-subsets of the $\underline{\mathbf{a}} \in \underline{\alpha}$ are exactly the supersets of the $\mathbf{a} \in \alpha$, i.e.

$$\{\mathbf{b} \in [n] : \neg\exists \underline{\mathbf{a}} \in \underline{\alpha} : \mathbf{b} \subseteq \underline{\mathbf{a}}\} = \{\mathbf{b} \in [n] : \exists \mathbf{a} \in \alpha : \mathbf{b} \supseteq \mathbf{a}\} \tag{5.53}$$

Suppose $\mathbf{b} \in [n]$ is an element of the LHS so that $\neg\exists\underline{\mathbf{a}} \in \underline{\alpha} : \mathbf{b} \subseteq \underline{\mathbf{a}}$ and assume that it is not contained in the RHS so that $\mathbf{b}$ is a non-superset of the $\mathbf{a} \in \alpha$. But then $\mathbf{b}$ must be a subset of some $\underline{\mathbf{a}} \in \underline{\alpha}$ since these are the maximal non-supersets of the $\mathbf{a} \in \alpha$. This contradicts our initial assumption. Hence, if $\mathbf{b}$ is in the LHS it must be in the RHS.

Now suppose that $\mathbf{b} \in [n]$ is an element of the RHS, i.e. it is a superset of some $\mathbf{a} \in \alpha$ and assume that it is not in the LHS because $\mathbf{b}$ is a subset of some $\underline{\mathbf{a}} \in \underline{\alpha}$. But then, since the $\underline{\mathbf{a}} \in \underline{\alpha}$ are the maximal non-supersets of the $\mathbf{a} \in \alpha$, $\mathbf{b}$ must be a non-superset of the $\mathbf{a} \in \alpha$ as well. Again this contradicts our initial assumption so that if $\mathbf{b}$ is in the RHS it must be in the LHS.

Therefore we have,

$$\overline{(\underline{\alpha})} = \min\{\mathbf{b} \in [n] : \neg\exists\underline{\mathbf{a}} \in \underline{\alpha} : \mathbf{b} \subseteq \underline{\mathbf{a}}\} = \min\{\mathbf{b} \in [n] : \exists\mathbf{a} \in \alpha : \mathbf{b} \supseteq \mathbf{a}\} = \alpha \quad (5.54)$$

5.9 Author contributions

AG developed the conceptual framework and ideas presented in this paper. Conducted the research and wrote the initial draft of the manuscript. AM rigorously verified the mathematical formulations and calculations presented in the manuscript. Also provided constructive feedback on draft revisions, figures, and other elements of the paper. MW provided critical feedback and suggestions for improving the manuscript, including revisions to figures and other content.

6

Significant subgraph mining for neural network inference with multiple comparisons correction

Aaron J. Gutknecht [1], Michael Wibral[1]

1 Campus Institute for Dynamics of Biological Networks, Georg-August University, Goettingen, Germany

Published as: ***Gutknecht, A. J., & Wibral, M. (2023). Significant subgraph mining for neural network inference with multiple comparisons correction. Network Neuroscience, 7(2), 389-410.***

Abstract

We describe how the recently introduced method of significant subgraph mining can be employed as a useful tool in neural network comparison. It is applicable whenever the goal is to compare two sets of unweighted graphs and to determine differences in the processes that generate them. We provide an extension of the method to dependent graph generating processes as they occur for example in within-subject experimental designs. Furthermore, we present an extensive investigation of the error-statistical properties of the method in simulation using Erdős-Rényi models and in empirical data in order to derive practical recommendations for the application of subgraph mining in neuroscience. In particular, we perform an empirical power analysis for transfer entropy networks inferred from resting state MEG data comparing autism spectrum patients with neurotypical controls. Finally, we provide a python implementation as part of the openly available IDTxl toolbox.

Author Summary

A key objective of neuroscientifc research is to determine how different parts of the brain are connected. The end result of such investigations is always a graph consisting of nodes corresponding to brain regions or nerve cells and edges between the nodes indicating if they are connected or not. The connections may be structural (an actual anatomical connection) but can also be functional – meaning that there is a statistical dependency between the activity in one part of the brain and the activity in another. A prime example of the latter type of connection would be the information flow between brain areas. Differences in the patterns of connectivity are likely to be responsible for and indicative of various neurological disorders such as autism spectrum disorders. It is therefore important that efficient methods to detect such differences are available. The key problem in developing methods for comparing patterns of connectivity is that there is generally a vast number of different patterns (it can easily exceed the number of stars in the milky way). In this paper we describe how the recently developed method of significant subgraph mining accounts for this problem and how it can be usefully employed in neuroscientific research.

6.1 Introduction

Comparing networks observed under two or more different conditions is a pervasive problem in network science in general, and especially in neuroscience. A fundamental question in these cases is if the observed patterns or motifs in two samples of networks differ solely due to chance or because of a genuine difference between the conditions under investigation. For example, a researcher may ask if a certain pattern of functional connections in a brain network reconstructed from magnetoencephalography (MEG) data is more likely to occur in individuals with autism spectrum disorder than in neurotypic controls, or whether an observed difference in occurrence is solely due to chance. What makes this question difficult to answer is the fact that the number of possible patterns in the network scales as 2^{l^2}, with l the number of network nodes, – leading to a formidable multiple comparison problem. Correcting for multiple comparisons with standard methods (e.g. Bonferroni) typically leads to an enormous loss of power as these methods do not exploit the particular properties of the network comparison problem.

By contrast, the recently developed Significant Subgraph Mining approach [38, 136] efficiently solves the network-comparison problem while maintaining strict bounds on type I error rates for between unit of observation designs. Within the landscape

of graph theoretic methods in neuroscience the distinguishing features of subgraph mining are, first, that it works with binary graphs, second, that it does not rely on summary statistics such as average clustering, modularity, or degree distribution (for review see for instance [137]), and third, that it is concerned with the *statistical* differences between graph generating processes rather than the distance between two individual graphs (for examples of such graph metrics see [138–140]). Subgraph mining can be considered the most fine-grained method possible for the comparison of binary networks in that it is in principle able to detect *any* statistical difference.

Here we describe how to adapt this method to the purposes of network neuroscience and provide a detailed study of it's error-statistical properties (family-wise error rate and statistical power) in both simulation and empirical data. In particular, we describe an extension of subgraph mining for within unit of observation designs that was, to our best knowledge, not described in the literature before. Furthermore, we utilize Erdős-Rényi networks as well as an empirical data set of transfer entropy networks to investigate the behaviour of the method under different network sizes, sample sizes, and connectivity patterns. Based on these analyses we discuss practical recommendations for the application of subgraph mining in neuroscience. Finally, we provide an openly available implementation of subgraph mining as part of the python toolbox IDTxl (`http://github.com/pwollstadt/IDTxl` [103]). The implementation readily deals with various different data structures encountered in neuroscientific research. These include directed and undirected graphs, between and within subject designs, as well as data with or without a temporal structure.

In the following section, we will explain the core ideas behind the original subgraph mining method as introduced in [38, 136] putting an emphasis on concepts and intuition, but also providing a rigorous mathematical exposition for reference. We then turn to the extension for within-subject designs before presenting the simulation-based and empirical investigation of subgraph mining.

6.2 Background and Theory: The original Subgraph Mining Method

Neural networks can usefully be described as graphs consisting of a set of nodes and a set of edges connecting the nodes ([137]). The nodes represent specific parts of the network such as individual neurons, clusters of neurons, or larger brain regions, whereas the edges represent relationships between these parts. Depending on whether the relationship of interest is symmetric (such as correlation) or asymmetric (such as Transfer Entropy or Granger Causality) the network can be modelled as an undirected or as a directed graph respectively. Once we have a decided upon an appropriate graph theoretic description, we can apply it to networks measured in two different experimental groups, resulting in two sets of graphs. In doing so, we are essentially sampling from two independent **graph-generating processes** (see Figure 6.1 for illustration).

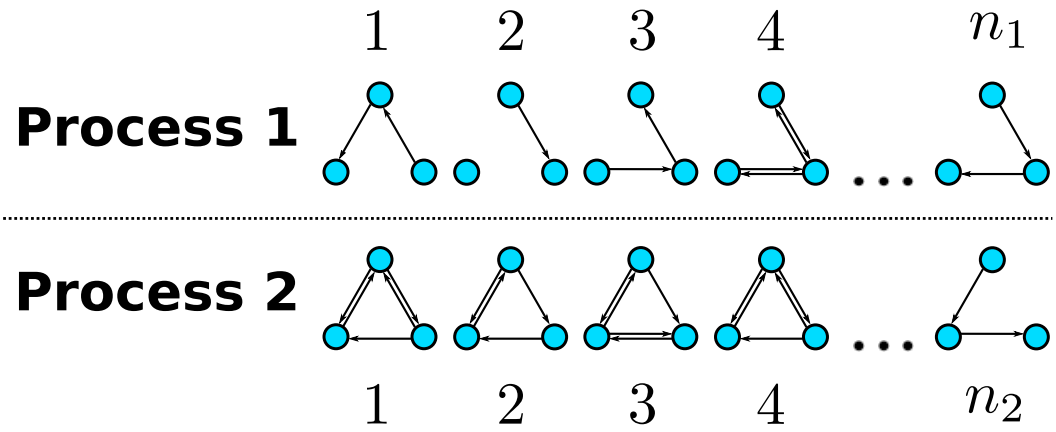

Fig. 6.1: Illustration of two graph-generating processes. Each process consists of randomly sampling individuals from a specific population and describing the neural activity of these individuals as a graph. The population underlying process 1 is sampled n_1 times and the population underlying process 2 is sampled n_2 times. The nodes may correspond to different brain areas while the edges describe any directed relationship between brain areas such as information transfer.

The key question is now if there are any significant differences between these two sets. However, since graphs are complex objects it is not immediately obvious how they should be compared. In principle, one may imagine numerous different possibilities. For instance, comparing the average number of connections of a node or the average number of steps it takes to get from one node to another. Instead of relying on such summary statistics, however, one may also take a more fine-grained approach by looking for differences with respect to any possible pattern, or more technically **subgraph**, that may have been observed. Does a particular edge occur

significantly more often in one group than in the other? What about particular bi-directional connections? Or are there even more complex subgraphs -consisting of many links- that are more frequently observed in one of the groups? Answering such questions affords a particularly detailed description of the differences between the two processes. Figure 6.2 shows examples of different subgraphs of a graph with three edges.

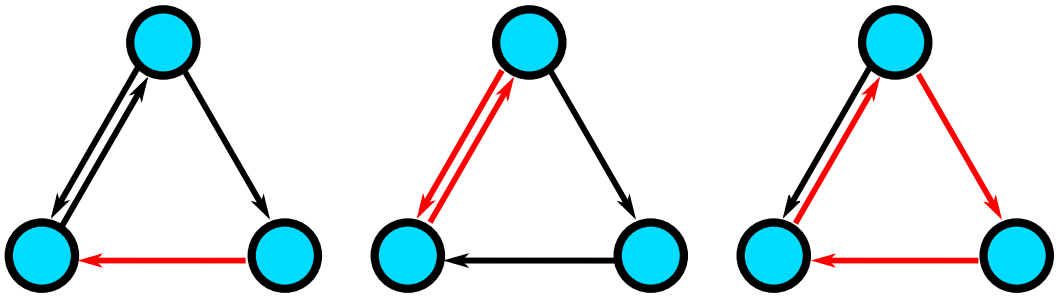

Fig. 6.2: Illustration of subgraphs with one edge (left), two edges (middle), and three edges (right) of a graph with three nodes.

The process of enumerating all subgraphs for which there is a significant difference between the groups is called **significant subgraph mining** [136]. The goal is to identify all subgraphs that are generated with different probabilities by the two processes. The main difficulty underlying significant subgraph mining is that the number of possible subgraphs grows extremely quickly with the number of nodes. For a directed graph with seven nodes, it is already in the order of 10^{14}. This not only imposes runtime constraints but also leads to a severe multiple comparisons problem. Performing a significance test for each potential subgraph and then adjusting by the number of tests is not a viable option because the resulting test will have an extremely poor statistical power. As will be detailed later, due to the discreteness of the problem the power may even be exactly zero because p-values low enough to reach significance can in principle not be achieved. In the following sections we will describe the original (between-subjects) significant subgraph mining method developed by [38, 136] by first setting up an appropriate probabilistic model, explaining how to construct a significance test for a particular subgraph, and finally, detailing two methods for solving the multiple comparisons problem.

Probabilistic Model

We are considering two independently sampled sets of directed graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ describing, for instance, connections between brain regions in two experimental groups. Each set contains one graph per subject in the corresponding group and

we assume that the (fixed) sample sizes of each group are $n_1 = |\mathcal{G}_1|$ and $n_2 = |\mathcal{G}_2|$. All graphs are defined on the same set of nodes $V = \{1, 2, ..., l\}$ but may include different sets of links (edges) $E \subseteq V \times V$. The graphs are assumed to have been generated by two potentially different graph-generating processes. Each process can be described by a random $l \times l$ adjacency matrix of, possibly dependent, Bernoulli random variables:

$$\mathbf{X}^{(k)} = \begin{bmatrix} X_{11}^{(k)} & X_{12}^{(k)} & ... & X_{1l}^{(k)} \\ X_{21}^{(k)} & X_{22}^{(k)} & ... & X_{1l}^{(k)} \\ ... & ... & ... & ... \\ X_{l1}^{(k)} & ... & ... & X_{ll}^{(k)} \end{bmatrix} \tag{6.1}$$

where the superscript $k = 1, 2$ indicates the group and

$$X_{ij}^{(k)} \sim Bern(p_{ij}^{(k)}), \;\; 1 \leqslant i, j \leqslant l \tag{6.2}$$

Each of those variables tells us whether the corresponding link from node i to node j is present ("1") or absent ("0"). A graph-generating process can be fully characterized by the probabilities with which it generates possible subgraphs. Specifically, there is one such probability for each possible graph $G = (V, E_G)$ on the nodes under consideration. The probability that G occurs as a subgraph of the generated graph in group k is given by

$$\pi_G^{(k)} = \mathbb{P}\left(\bigcap_{(i,j) \in E_G} \{X_{ij}^{(k)} = 1\} \right) \tag{6.3}$$

where (i, j) indicates an individual link from node i to node j. It is important to note that $\pi_G^{(k)}$ denotes the the probability that all the edges of G are realized *plus possibly some additional edges*. This is to be distinguished from the probability that *exactly* the graph G is realized. In the following we will always refer to the probability $\pi_G^{(k)}$ as the **subgraph probability** of G. A graph generating process is completely specified when all it's subgraph probabilities are specified. So to sum up, we can model the two sets of directed graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ as realizations of two independent graph generating processes $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. Process $\mathbf{X}^{(1)}$ generates graphs according to subgraph probabilities $\pi_G^{(1)}$ whereas the subgraph probabilities for process $\mathbf{X}^{(2)}$ are given by $\pi_G^{(2)}$. Based on this probabilistic model we may now proceed to test for differences between the two processes.

Testing Individual Subgraphs

Our goal now is to find those subgraphs G that are generated with different probabilities by the two processes. If the two processes describe two distinct experimental groups, this means that we are trying to identify subgraphs whose occurrence depends on group membership. Thus, for each possible subgraph G, we are testing the null hypothesis of *equal subgraph probabilities*, or equivalently, of *independence of subgraph occurrence from group membership*

$$H_0^G : \pi_G^{(1)} = \pi_G^{(2)} \tag{6.4}$$

against the alternative of unequal subgraph probabilities or dependence on group membership

$$H_1^G : \pi_G^{(1)} \neq \pi_G^{(2)} \tag{6.5}$$

In order to test such a null-hypothesis we have to compare how often the subgraph G occurred in each group and determine if the observed difference could have occurred by chance, i.e. if the probability of such a difference would be larger than the significance level α under the null-hypothesis. The relevant data for this test can be summarized in a 2×2 contingency table:

Subgraph G	Occurrences	Non-Occurrences	Total
Group 1	$f_1(G)$	$n_1 - f_1(G)$	n_1
Group 2	$f_2(G)$	$n_2 - f_2(G)$	n_2
Total	$f(G)$	$n - f(G)$	n

where $f_i(G)$ denotes the *observed* absolute frequency of subgraph G in Group i, $f(G) = f_1(G) + f_2(G)$ denotes the *observed* absolute frequency of G in the entire data set, and $n = n_1 + n_2$ is the total sample size. In the following, we will use $F_i(G)$ and $F(G)$ to denote the corresponding *random* absolute frequencies. Given our model assumptions above, the numbers of occurrences in each group are independent Binomial variables: On each of the n_1 (or n_2) independent trials there is a fixed probability $\pi_G^{(1)}$ (or $\pi_G^{(2)}$) that the subgraph G occurs. This means that our goal is to compare two independent Binomial proportions. This can be achieved by utilizing Fisher's exact test [38, 136] which has the advantage that it does not require any minimum number of observations per cell in the contingency table.

The key idea underlying Fisher's exact test is to condition on the total number of occurrences $f(G)$. Specifically, the random variable $F_1(G)$ can be shown to follow a hypergeometric distribution under the null-hypothesis and conditional on the total

number of occurrences. In other words, if the null-hypothesis is true and given the total number of occurrences, the n_1 occurrences and non-occurrences of subgraph G in Group 1 are assigned as if they were drawn randomly without replacement out of an urn containing exactly $f(G)$ occurrences and $n - f(G)$ non-occurrences (see Figure 6.3). $F_1(G)$ can now be used as a test-statistic for the hypothesis test.

Since we are interested in differences between the graph generating processes in either direction the appropriate test is a *two-sided* one. For a right-sided test of the null-hypothesis against the alternative $\pi_G^{(1)} > \pi_G^{(2)}$ the p-value can be computed as

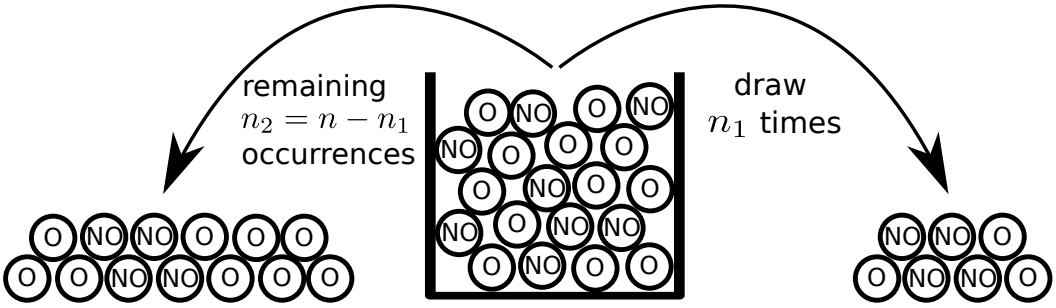

Fig. 6.3: Comparing two Binomial proportions using Fisher's exact test. Under the null-hypothesis and conditional on the total number of occurrences of a subgraph, the occurrences are distributed over the groups as if drawn at random without replacement out of an urn containing one ball per subject. The balls are labelled 'O' if the subgraph occurred in the corresponding subject and 'NO' if it did not occur. In the illustration $n = 20$ (number of total measurements, balls), $n_1 = 7$ (number of measurements for group 1, black balls), and $f(G) = 12$ (number of occurences, balls with 'O'). The seven balls drawn for group 1 are shown to the right of the urn. They include three occurrences and four non-occurrences. This result would lead to an insignificant p-value of ≈ 0.5

$$p_G^R = \sum_{k=f_1(G)}^{min(f(G),n_1)} hyp(k; N, f(G), n_1) \tag{6.6}$$

summing up the probabilities of all possible values of $f_1(G)$ *larger than or equal to* the one actually observed. Note that $f_1(G)$ cannot be larger than $min(f(G), n_1)$ because the number of occurrences in Group 1 can neither be larger than the sample size n_1 nor larger than the total number of occurrences $f(G)$. A left-sided p-value can be constructed analogously. The two-sided test rejects the null-hypothesis just in case the two-sided p-value

$$p_G = 2 * min(p_G^L, p_G^R) \tag{6.7}$$

is smaller than or equal to α .

Multiple Comparisons

Since there may be a very large number of possible subgraphs to be tested we are faced with a difficult multiple comparisons problem. For a directed graph with 7 nodes the number of possible subgraphs is already in the order of 10^{14}. If we were to use this number as a Bonferroni correction factor the testing procedure would have an exceedingly low statistical power meaning that it would be almost impossible to detect existing differences in subgraph probabilities. In the following, we will describe two methods for solving the multiple comparisons problem: the Tarone correction [141] and the Westfall-Young permutation procedure [142] which have been used in the original exposition of significant subgraph mining by [38, 136].

Tarone's Correction

The subgraph mining problem is discrete in the sense that there is only a finite number of possible p-values. This fact can be exploited to drastically reduce the correction factor. The key insight underlying the Tarone correction is that given any total frequency $f(G)$ of a particular subgraph G there is a *minimum achievable p-value* which we will denote by p^*_G. Intuitively, this minimum achievable p-value is reached if the $f(G)$ occurrences are distributed as unevenly as possible over the two groups. We may now introduce the notion of the set $T(k)$ of $\frac{\alpha}{k}$-**testable subgraphs**:

$$T(k) = \{G \subseteq G_C : p^*_G \leqslant \frac{\alpha}{k}\} \tag{6.8}$$

containing all subgraphs whose minimum achievable p-value is smaller than or equal to $\frac{\alpha}{k}$. Following Tarone, the number of elements of this set can be denoted by $m(k) = |T(k)|$. Tarone et al then showed that the smallest integer k such that $\frac{m(k)}{k} \leqslant 1$ is a valid correction factor in the sense that the probability of rejecting a true null-hypothesis, the **family-wise error rate (FWER)** , is bounded by α [141]. Moreover, the family-wise error rate is controlled no matter which or how many null-hypotheses are true (see Supporting Information for proof). This property is called **strong control**. A slight improvement of this correction factor was proposed by [143] (for details see Supporting Information). Figure 6.4 illustrates the concepts of testable, untestable, and significant subgraphs.

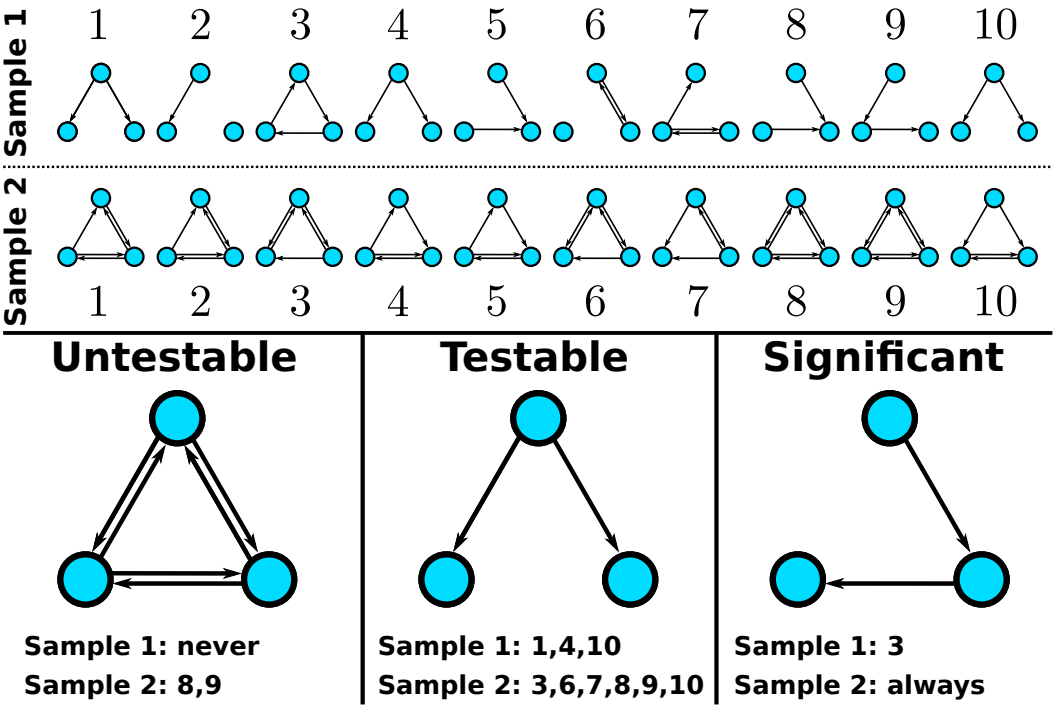

Fig. 6.4: Examples of 0.05-untestable, 0.05-testable, and significant subgraphs for a data set consisting of 10 graphs per group (top panel). The fully connected graph is untestable at level 0.05 because it only occurs twice in the data set (group 2 samples 8 and 9) leading to a minimum achievable p-value of ≈ 0.47. The graph shown on the bottom middle is testable at level 0.05 since it occurs 9 times in total. This means that its minimum achievable p-value is ≈ 0.0001. However, it is not significant with an actual (uncorrected) p-value of ≈ 0.37. The graph shown on the bottom right reaches significance using Tarone's correction factor $K(0.05) = 17$. It occurs every time in group 2 but only once it group 1 which results in a corrected p-value of ≈ 0.02.

Westfall-Young Correction

The familiy-wise error rate with respect to a corrected significance level δ can be expressed in terms of the cumulative distribution function of the smallest p-value associated with a true null-hypothesis: the event that there is at least one false positive is identical with the event that the smallest p-value associated with a true null-hypothesis is smaller than δ. The same applies to the *conditional* family-wise error rate given the total occurrences of each graph in the data set:

$$CFWER(\delta) = \mathbb{P}\left(\min_{G \in \mathcal{G}_0}(P_G) \leqslant \delta | F = f\right) \tag{6.9}$$

where $\mathcal{G}_0$ is the set of subgraphs for which the null is true and F is the vector of the total occurrences of each subgraph. This means that if the correction factor is chosen

as the α-quantile of the distribution in 6.9 the family-wise error rate is controlled. The problem is that we cannot evaluate the required distribution because we don't know which hypotheses are true. The idea underlying the Westfall-Young correction is to instead define the correction factor as the α-quantile of the distribution of the minimal p-value across *all* subgraphs and under the *complete* null-hypothesis (stating that all null hypotheses are true). This correction factor always provides **weak control** of the FWER in the sense that the FWER is bounded by α under the complete null-hypothesis (the issue of strong control is addressed in the Discussion section). It can be estimated via permutation strategies. The procedure is as follows: First, we may represent the entire data set by the following table

Subject	Group	G_1	G_2	...	G_m
1	0	0	1	...	1
2	0	1	1	...	1
...	...	...	...	...	...
n_1	0	0	0	...	1
$n_1 + 1$	1	1	1	...	1
...	...	...	...	...	...
$n_1 + n_2$	1	0	1	...	1

The columns labelled G_i tell us if subgraph G_i was present or absent in the different subjects (rows). The column labelled "Group" describes which group the different subjects belong to. Under the complete null-hypothesis the group labels are arbitrarily exchangeable. This is because, given our independence assumptions, all the observed graphs in the data set are independent and identically distributed samples from the same underlying distribution in the complete null-case. The column of group labels is now shuffled, reassigning the graphs in the data set to the two groups. Based on this permuted data set we can recompute a p-value for each G_i and determine the smallest of those p-values. Repeating this process many times allows us to obtain a good estimate of the distribution of the smallest p-value under the complete null-hypothesis. The Westfall-Young correction factor is then chosen as the α-quantile of this permutation distribution. Since the number of permutations grows very quickly with the total sample size, it is usually not possible to evaluate all permutations. Instead, one has to consider a much smaller random sample of permutations in order to obtain an approximation to the permutation distribution. This procedure can be shown to be valid as long as the identity permutation (i.e. the original data set) is always included [144].

This concludes our discussion of the original subgraph mining method. Figure 6.5 provides a schematic illustration of the essential steps. In the next section,

we describe how the method can be extended to be applicable to within-subject experimental designs which abound in neuroscience.

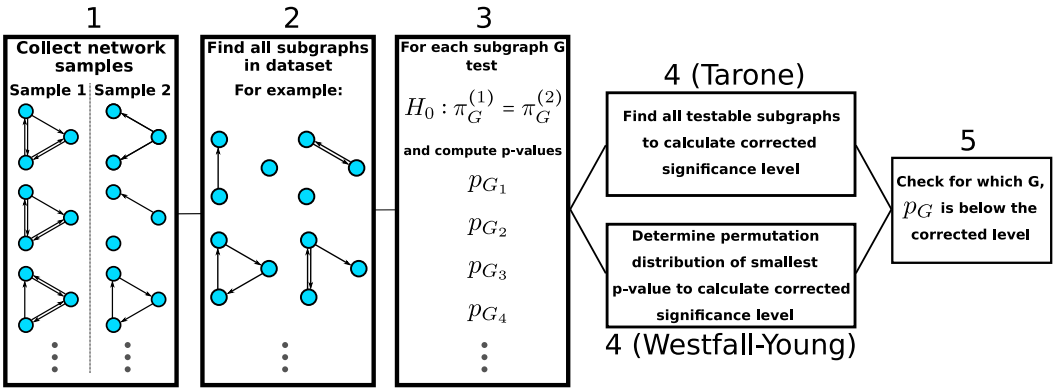

Fig. 6.5: Schematic illustration of significant subgraph mining. Note that for computational efficiency various shortcuts can be employed. The figure describes conceptually how significant subgraph mining works rather than it's fastest possible implementation (see for example [38] for a fast algorithm implementing the Westfall-Young correction).

6.3 Extension to Within-Subject Designs

So far we have considered networks associated with subjects from two groups and we assumed that the numbers of occurrences of a subgraph in the two groups are independent of each other. However, there are many cases in which there is only a single group of subjects and we are interested in how the networks differ between two experimental conditions. Since the same subjects are measured in both conditions, the independence assumption is not warranted anymore. Because Fisher's exact test assumes independence, the approach described above has to be modified. In particular, in case of dependence, the null-distribution of the number of occurrences in the first group / condition will in general not be a hypergeometric distribution potentially leading to inflated type I error rates in Fisher's exact test. An appropriate alternative is McNemars test for marginal homogeneity. It essentially tests the same null-hypothesis as Fisher's exact test, but is based on a wider probabilistic model of the graph generating processes. In particular, the independence assumption is relaxed allowing for dependencies between the two experimental conditions: Whether a subgraph occurs in condition A in a particular subject may affect the probability of its occurrence in condition B and vice versa. Suppose we are observing n subjects in two conditions. We may denote the random adjacency matrices corresponding the i-th subject in condition 1 and 2 by $\mathbf{X}_i^{(1)}$ and $\mathbf{X}_i^{(2)}$, respectively. Then the probabilistic model for the graph-generating processes is:

$$\begin{pmatrix} \mathbf{X}_1^{(1)} \\ \mathbf{X}_1^{(2)} \end{pmatrix}, \begin{pmatrix} \mathbf{X}_2^{(1)} \\ \mathbf{X}_2^{(2)} \end{pmatrix}, \ \ldots \ , \begin{pmatrix} \mathbf{X}_n^{(1)} \\ \mathbf{X}_n^{(2)} \end{pmatrix} \text{ i.i.d.} \tag{6.10}$$

For each subject there is an independent and identically distributed realization of the two graph-generating processes. The two processes themselves may be dependent since they describe the same subject being observed under two conditions. The distributions of $\mathbf{X}_i^{(1)}$ and $\mathbf{X}_i^{(2)}$ are again determined by the subgraph probabilities $\pi_G^{(1)}$ and $\pi_G^{(2)}$ and for any particular G we would like to test the null-hypothesis:

$$H_0^G : \pi_G^{(1)} = \pi_G^{(2)} \tag{6.11}$$

The idea underlying McNemar's test is to divide the possible outcomes for each subject into four different categories: 1) G occurred in both conditions, 2) G occurred in neither condition, 3) G occurred in condition 1 but not in condition 2, 4) G occurred in condition 2 but not in condition 1. The first two categories are called *concordant pairs* and the latter two are called *discordant pairs*. The discordant pairs are of particular interest because differences in subgraph probabilities between the

two conditions will manifest themselves in the relative number of the two types of discordant pairs: If $\pi_G^{(1)} > \pi_G^{(2)}$, then we would expect to observe the outcome 'G occurred only in condition 1' more frequently than the outcome 'G occurred only in condition 2'. Conversely, if $\pi_G^{(2)} > \pi_G^{(1)}$, than we would expect to observe the latter type of discordant pair more frequently. The frequency of any of the four categories can be represented in a contingency table:

Condition 1 / Condition 2	Yes	No	Total
Yes	Y_{11}^G	Y_{10}^G	$F_1(G)$
No	Y_{01}^G	Y_{00}^G	n - $F_1(G)$
Total	$F_2(G)$	n-$F_2(G)$	n

The variables $Y_{11}^G, Y_{11}^G, Y_{21}^G, Y_{22}^G$ are the counts of the four categories. The numbers of occurrences in each condition $F_1(G)$ and $F_2(G)$ appear in the margins of the contingency table. McNemar's test uses the upper right entry, Y_{10}^G, as the test-statistic. Conditional on the total number of discordant pairs, $Y_{10}^G + Y_{01}^G$, and under the null-hypothesis, this test-statistic has a binomial distribution

$$Y_{10}^G \mid Y_{10}^G + Y_{01}^G = d \overset{H_0}{\sim} Bin\left(d, \frac{1}{2}\right) \tag{6.12}$$

If there are exactly d discordant pairs and the probability of G is equal in both conditions, then both types of discordant pairs ('only in condition 1' or 'only in condition 2') occur independently with equal probabilities in each of the d subjects where a discordant pair was observed. A two-sided test can be constructed in just the same way as described above for the between-subjects case. First, we construct right- and left-sided p-values as:

$$p_G^L = \sum_{k=0}^{y_{10}^G} Bin\left(k; d, \frac{1}{2}\right) \qquad p_G^R = \sum_{k=y_{10}^G}^{d} Bin\left(k; d, \frac{1}{2}\right) \tag{6.13}$$

Then the two-sided p-value is

$$p_G = 2 * \min(p_G^L, p_G^R) \tag{6.14}$$

Exactly like the Fisher's test, McNemar's test also has a minimal achievable p-value. The only difference is that it is not a function of the total number of occurrences in condition A, but a function of the number of discordant pairs. The Tarone correction described above remains valid if Fisher's exact test is simply replaced by McNemar's test. The Westfall-Young procedure requires some modifications because

the permutation strategy described above is not valid anymore. The problem is that, because of possible dependencies between the conditions, condition labels are not arbitrarily exchangeable under the complete null-hypothesis. Instead we have to take a more restricted approach and only exchange condition labels *within subjects*. In doing so, we are not only keeping the total number of occurrences $F(G)$ constant for each subgraph, but also the total number of discordant pairs $D(G)$. Accordingly, the theoretical Westfall-Young correction factor, estimated by the modified permutation strategy, is the α-quantile of the conditional distribution of the smallest p-value given $F = f$ and $D = d$ and under the complete null-hypothesis (where F and D are the vectors of total occurrences and discordant pair counts for all subgraphs).

6.4 Validation of Multiple Comparisons Correction Methods using Erdős-Rényi Models

In this section we empirically investigate the family-wise error rate and statistical power of the multiple comparison correction methods for significant subgraph mining described above. In doing so we will utilize Erdős-Rényi models for generating random graphs. In these models the edges occurs independently with some common probability p_i in each graph-generating process. This means that the subgraph probability for a graph $G = (V, E_G)$ in process i is p_i raised to the number of edges G consists of:

$$\pi_G^{(i)} = p_i^{|E_G|} \tag{6.15}$$

If p_i is the same for both graph-generating processes ($p_1 = p_2$), then the complete null-hypothesis is satisfied. By contrast, if p is chosen differently for the two processes ($p_1 \neq p_2$), then the null-hypothesis of equal subgraph probabilities is violated for all subgraphs, i.e. the *complete alternative* is satisfied. We used the former setting for the purposes of FWER estimation and the latter for power analysis. Furthermore, the two graph-generating processes were simulated independently of each other which corresponds to the between-subjects case. Accordingly, Fisher's exact test was used throughout.

Family-Wise Error Rate

In order to empirically ascertain that the desired bound on the family-wise error rate is maintained by the Tarone and Westfall-Young corrections in the subgraph mining context, we performed a simulation study based on Erdős-Rényi models. We tested sample sizes $n = 20, 30, 40$, network sizes $l = 2, 4, 6, 8, 10$, and connection densities p = 0.1, 0.2, 0.3. For each combination of these values we carried out 1000 simulations and estimated the empirical FWER as the proportion of simulations in which one or more significant subgraphs were identified. Figure 6.6 shows the results of this analysis. The FWER is below the prespecified $\alpha = 0.05$ in all cases for the Tarone and Bonferroni corrections and always within one standard error of this value for the Westfall-Young correction. The Bonferroni correction is most conservative. In fact, the FWER quickly drops to exactly zero since the Bonferroni-corrected level is smaller than the smallest possible p-values. The Tarone-correction reaches intermediate values of 0.1-0.3 while the Westfall-Young correction is always closest the prespecified level and sometimes even reaches it.

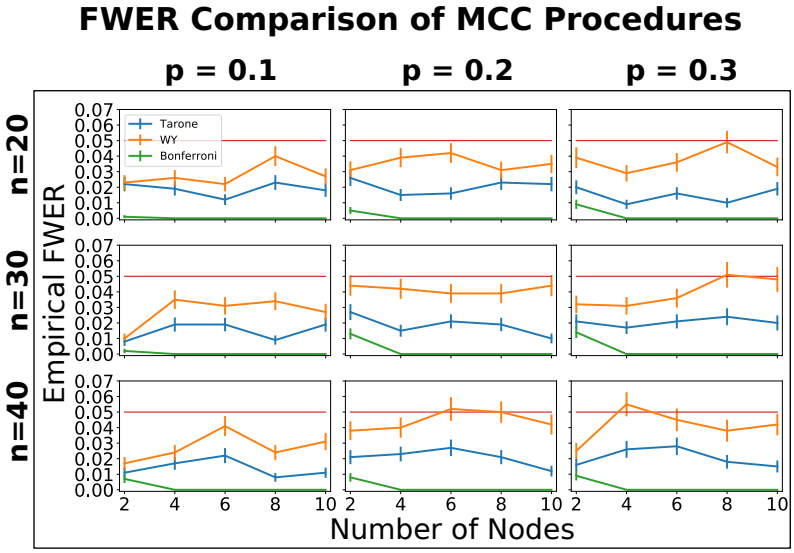

Fig. 6.6: Estimated family-wise error rates of Tarone, Westfall-Young, and Bonferroni corrections based on 1000 simulations and different sample sizes, connection densities, and network sizes. Error-bars represent one standard-error. The estimated FWER never exceeded the desired FWER of $\alpha = 0.05$ (red horizontal line) by more than one standard-error for all correction methods. In fact, it was always smaller than 0.05 except in three cases for the Westfall-Young correction (0.051, 0.052, and 0.055). The estimated FWERs of the three methods were always ordered in the same way: The Bonferroni correction had the smallest estimated FWER (at most 0.014), followed by the Tarone correction (at most 0.028), and the Westfall-Young correction (at most 0.055).

Power

We now turn our attention to the statistical power of the multiple comparison correction methods, i.e. their ability to detect existing differences between subgraph probabilities. Previous studies have used the empirical FWER as a proxy for statistical power [38, 136]. The rationale underlying this approach is that the more conservative a method is (i.e. the more the actual FWER falls below the desired significance level), the lower its statistical power. In the following we will take a more direct approach and evaluate the performance of the methods under the alternative hypothesis of unequal subgraph probabilities. Again we will utilize Erdős-Rényi models, only now with different connection densities $p_1 \neq p_2$ for the two graph-generating processes. The question is: How many subgraphs are we able to correctly identify as being generated with distinct probabilities by the two processes? The answer to this question will not only depend on the multiple comparisons correction used but also on the sample size, the network size, and the effect size. The effect size *for a particular subgraph* G can be identified with the magnitude of the difference of subgraph probabilities $|\pi_G^{(1)} - \pi_G^{(2)}|$. The larger this difference, the better the chances to detect the effect. In the following we will use the difference between the connection densities p_1 and p_2 as a measure of the effect size for the entire graph-generating processes.

In a simulation study we analyzed sample sizes n = 20, 30, 40. We set the probability of individual links for the first graph-generating process to $p_1 = 0.2$. The second process generated individual links with probability $p_2 = 0.2 + e$, where e = 0.1, 0.2, 0.3. Since p_1 and p_2 are chosen smaller than or equal to 0.5, the effect sizes for particular subgraphs are a decreasing function of the number of edges they consist of. In other words, the difference is more pronounced for subgraphs consisting only of few edges and can become very small for complex subgraphs. We considered network sizes $l = 2, 4, 6, 8, 10$. For each possible choice of n, e, and l we simulated 1000 data sets and applied significant subgraph mining with either Tarone, Westfall-Young or Bonferroni correction. The number of permutations for the Westfall-Young procedure was set to 10000 as recommended in previous studies [38]. The two graph-generating processes were sampled independently (between subjects case) and accordingly Fisher's exact test was utilized. The results are shown in Figure 6.7.

As expected the average number of detected significant subgraphs is an increasing function of both sample size and effect size. The relationship between detected differences and number of nodes is less straightforward. Generally, there is an increasing relationship, but there are a few exceptions. The likely explanation for

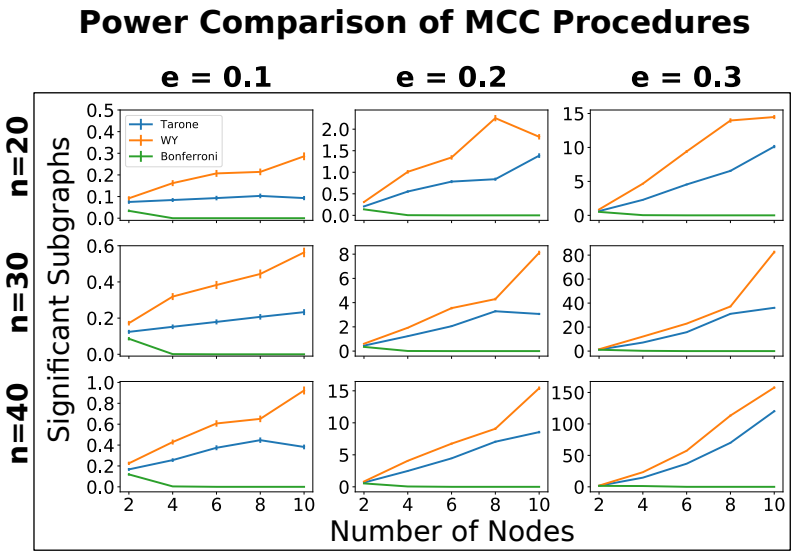

Fig. 6.7: Average number of significant subgraphs identified depending on correction method, samples size, network size, and effect size. Error bars represent one standard error. The number of identified subgraphs increases with sample size (rows) and effect size (columns) for all correction methods.

this phenomenon is that there is a trade-off between two effects: on the one hand, the larger the number of nodes the more differences there are to be detected. But on the other hand, the larger the number of nodes the more severe the multiple comparisons problem becomes which will negatively affect statistical power. For some parameter settings this latter effect appears to be dominant. The most powerful method is always the Westfall-Young correction followed by the Tarone correction. The Bonferroni correction has the worst performance and its power quickly drops to zero because the corrected threshold can in principle not be attained.

Generally, only a very small fraction of existing differences is detectable. Since the graphs are generated by independently selecting possible links with a fixed probability, the subgraph probability is a decreasing function of the number of links a subgraph consists of. Complex subgraphs are quite unlikely to occur and will therefore not be testable. Additionally, the difference between subgraph probabilities $\pi_G^{(1)}$ and $\pi_G^{(2)}$ decreases with increasing subgraph complexity making this difference more difficult to detect. For instance, if e = 0.3, then the difference in subgraph probabilities for subgraphs with 10 nodes is about 0.001. Accordingly, even with a sample size of 40, only a small fraction of existing differences is detectable.

Voxel ID	Location
0	Cerebellum
1	Cerebellum
2	Lingual Gyrus / Cerebellum
3	Posterior Cingulate Cortex (PCC)
4	Precuneus
5	Supramarginal Gyrus
6	Precuneus

Tab. 6.1: Voxel IDs and corresponding brain regions

6.5 Empirical Power Analysis with Transfer Entropy Networks

We applied the subgraph mining method to a data set of resting state MEG recordings comparing 20 autism spectrum disorder patients to 20 neurotypical controls. The details of the study are described in [40]. Here, seven voxels of interest were identified based on differences in local active information storage; subsequently timecourses of neural mass activity in these voxels were reconstructed by means of a linear constraint minimum variance (LCMV) beamformer. The locations of the voxels are shown in Table 6.1. We applied an iterative greedy method to identify multivariate **transfer entropy** networks on these voxels ([36, 145]). This is at present considered the best ([146]) means of measuring neural communication in data (also called "communication dynamics" [147]). The goal of this method is to find for each target voxel a set of source voxels such that 1) the total transfer entropy from the sources to the target is maximized, and 2) each source provides significant transfer entropy conditional on all other source voxels in the set. The outcome of this procedure is one directed graph per subject where each link represents significant information transfer from one voxel to another (conditional on the other sources). Accordingly, we are in a setting in which subgraph mining is applicable. The inferred transfer entropy graphs are shown in Figures 6.8, 6.9. Note that the edges are labeled by numbers that represent the time lags at which information transfer occurred. The parameters of the network inference algorithm were chosen so that lags are always multiples of five. Since the sampling rate was 1200Hz this corresponds to a lag increment of ≈ 4ms. So the graph representation also contains information about the temporal structure of information transfer and differences in this structure can be detected by subgraph mining as well. For example, even if the probability of detecting information transfer from voxel 0 to voxel 1 is the same in both groups, this transfer may be more likely to occur at a time lag of 5 ($\approx 4ms$) in

the autism group whereas it may be more likely to occur at a time lag of 10 ($\approx 8ms$) in the control group.

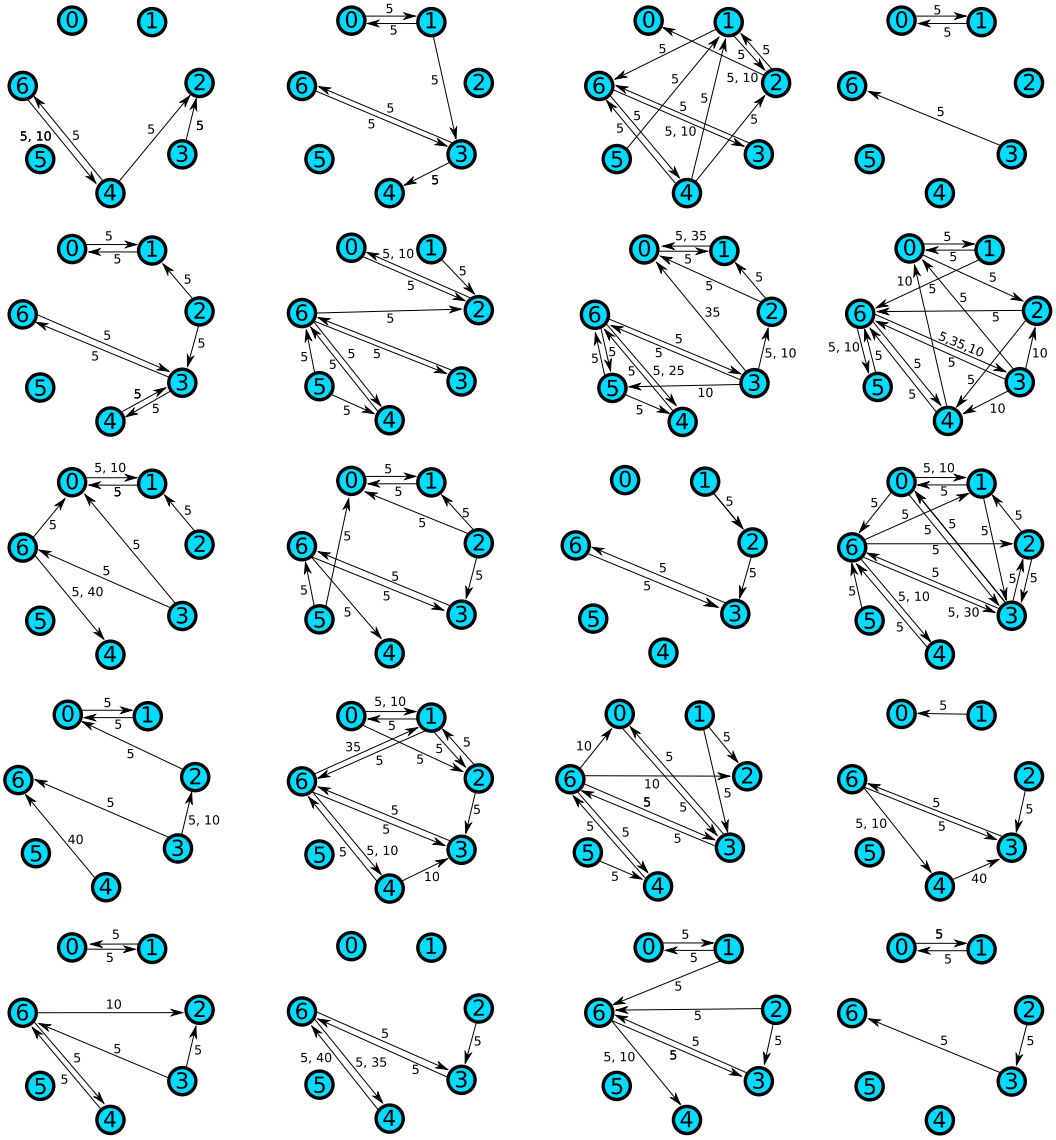

Fig. 6.8: Transfer Entropy networks detected in autism spectrum group.

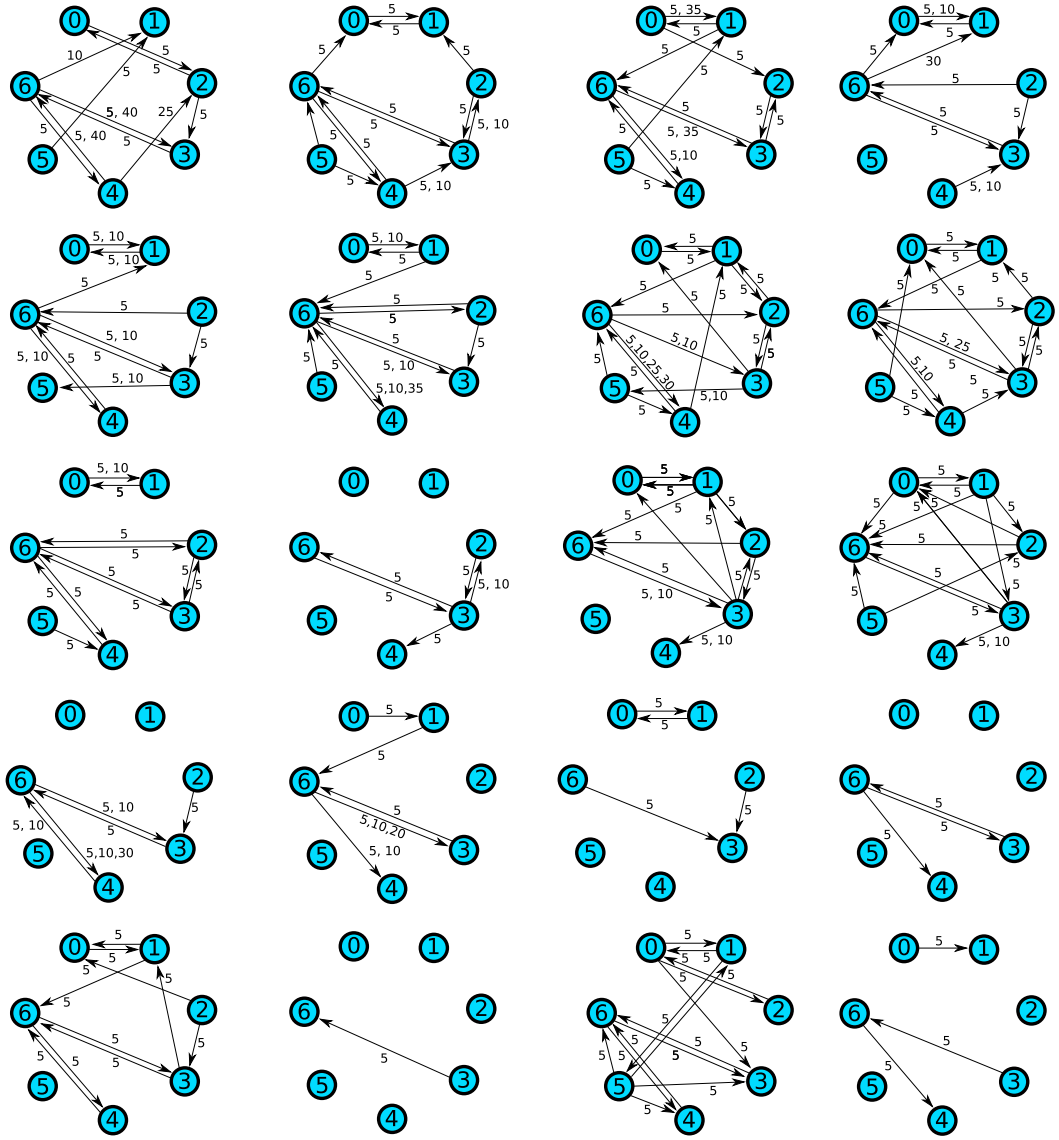

Fig. 6.9: Transfer Entropy networks detected in control group.

We applied subgraph mining with both Tarone and Westfall-Young correction to this data set. No significant differences between the ASD group and control group could be identified. Due to the rather small sample size, this result is not entirely unexpected. For this reason, we performed an empirical power analysis in order to obtain an estimate of how many subjects per group are required in order to detect existing differences between the groups. This estimate may serve as a useful guideline for future studies. The power analysis was performed in two ways: First, by resampling links independently using their empirical marginal frequencies, and second, by resampling from the empirical joint distribution, i.e. randomly drawing networks from the original data sets with replacement.

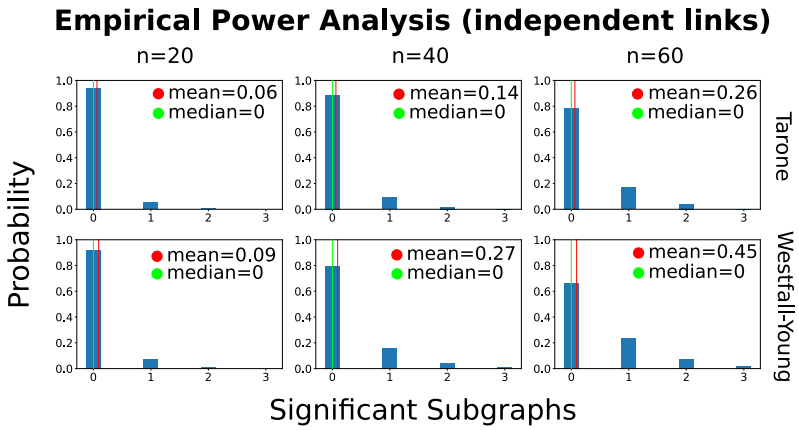

Fig. 6.10: Results of empirical power analysis assuming *independence* of links. We simulated sample sizes 20, 40, and 60 per group and carried out 400 simulations in each setting. The histograms describe the fractions of simulations in which different numbers of significant subgraphs were detected.

The results of the power analysis assuming independent links are shown in Figure 6.10. We simulated sample sizes 20, 40, and 60 per group and carried out 400 simulations for each setting. The first notable outcome is that the original data are strikingly different from the results seen in independent sampling of links. In particular, the number of testable graphs is far higher in the original data (1272) than in the independently resampled data (28.7 on average and 55 at most). This indicates strongly that the processes generating the networks in ASD patients as well as controls do not generate links independently. Rather, there seem to be dependencies between the links such that some links tend to occur together making it more likely that subgraphs consisting of these links will reach testability. Accordingly, in the case of independent resampling much larger sample sizes are needed in order to detect the differences between the groups. Even in the $n = 60$ per group setting there were only 0.26 (Tarone) and 0.45 (Westfall-Young) significant subgraphs on

average. There was no simulation in which more than three significant subgraphs were detected.

The simulation results of the empirical power analysis based on the empirical joint distribution are shown in Figure 6.11. Again we used sample sizes 20, 40 and 60. The average number of testable subgraphs is in the same order of magnitude as in the original data set for the $n = 20$ setting (≈ 5200). Moreover, the number of identified significant subgraphs is far greater than in independent sampling for all sample sizes. The Westfall-Young correction identifies more subgraphs on average than the Tarone correction: 17.41 compared to 0.86 for $n = 20$, 202.20 compared to 14.88 for $n = 40$, and 831,24 compared to 100.62 for $n = 60$. The distributions are always highly skewed with more probability mass on smaller values. This is reflected in the median values also shown in the figure. For example, notwithstanding the average value of 14.88 significant subgraphs in the $n = 40$ setting with Tarone correction, the empirical probability of not fining any significant subgraph is still $\approx 42\%$. For the Westfall-Young correction this probability is only $\approx 1.8\%$ in the $n = 40$ setting. In the $n = 60$ setting both methods have high empirical probability to detect significant differences. In fact, the Westfall-Young correction always found at least one difference and the Tarone correction only failed to find differences in 2.5% of simulations. The total number of detected differences can be in the thousands in this setting.

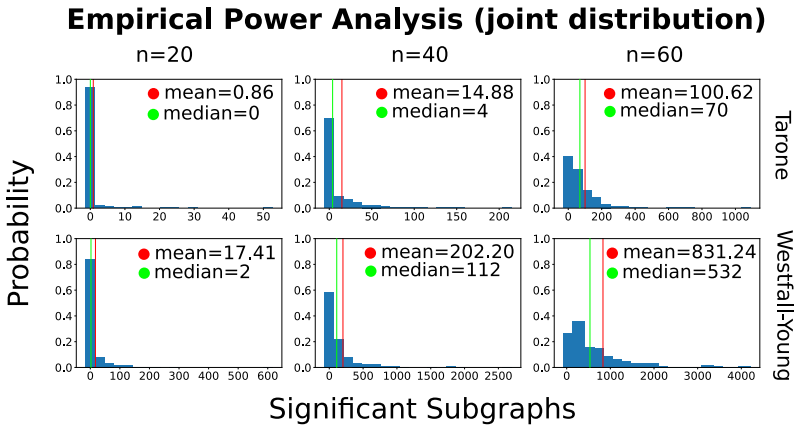

Fig. 6.11: Results of empirical power analysis performed by sampling from the empirical joint distribution. We simulated sample sizes 20, 40, and 60 per group and carried out 400 simulations in each setting. The histograms describe the fractions of simulations in which different numbers of significant subgraphs were detected.

Since in the $n = 60$ setting both methods are likely to detect some of the existing differences, we performed a subsequent analysis to narrow down the effect sizes that can be detected in this case. For each possible effect size (any multiple of 0.05 up to 0.35) we enumerated all subgraphs with this effect size and calculated

their empirical detection probabilities among the 400 simulations. In total there were about 3.7 million subgraphs occurring with different empirical probabilities in the two groups. Most of these (99.5%) are subgraphs that occur exactly once in the entire data set. One important reason for this phenomenon is the following: suppose a network contains a subgraph that occurs only once in the data set. Then removing any other edges or combination of edges from the network will again result in a subgraph that only occurs once in the data set. Consider for example the last network in the second row in Figure 6.8. It contains a connection from node 6 to node 3 at a lag of 35 samples. This connection does not occur in any other network. This means that if any combination of the other 18 links occurring in the network is removed, the result will again be a uniquely occurring subgraph. There are $2^{18} = 262144$ possibilities for doing so in this case alone.

The averages of the empirical detection probabilities for each effect size are shown in Figure 6.12 (upper plots). An interesting outcome is that the detection probability is not a strictly increasing function of the effect size. Rather there is a slight drop from effect sizes 0.25 to 0.3. Given the standard errors of the estimates this result might still be explained by statistical fluctuation (the two standard error intervals slightly overlap). However, in general this type of effect could also be real because the effect size is not the only factor determining detection probability. This is illustrated in Figure 6.12 (lower plots) which shows average detection probability over the smaller of the two occurrence probabilities $min(\pi_G^{(1}, \pi_G^{(2})$. It turns out that the more extreme this probability is, the more likely the effect is to be detected. The highest detection probability is found if the empirical probability of occurrence is zero in one of the groups. For this reason it can in fact be true that the detection probability is on average higher for effect sizes of size 0.25 than 0.3, if the absolute occurrence probabilities are more extreme in the former case. In the data analysed here this is in fact the case: roughly half of the subgraphs with effect size 0.25 do have occurrence probability zero in one of the groups whereas this is not true for any of the subgraphs with effect size 0.3.

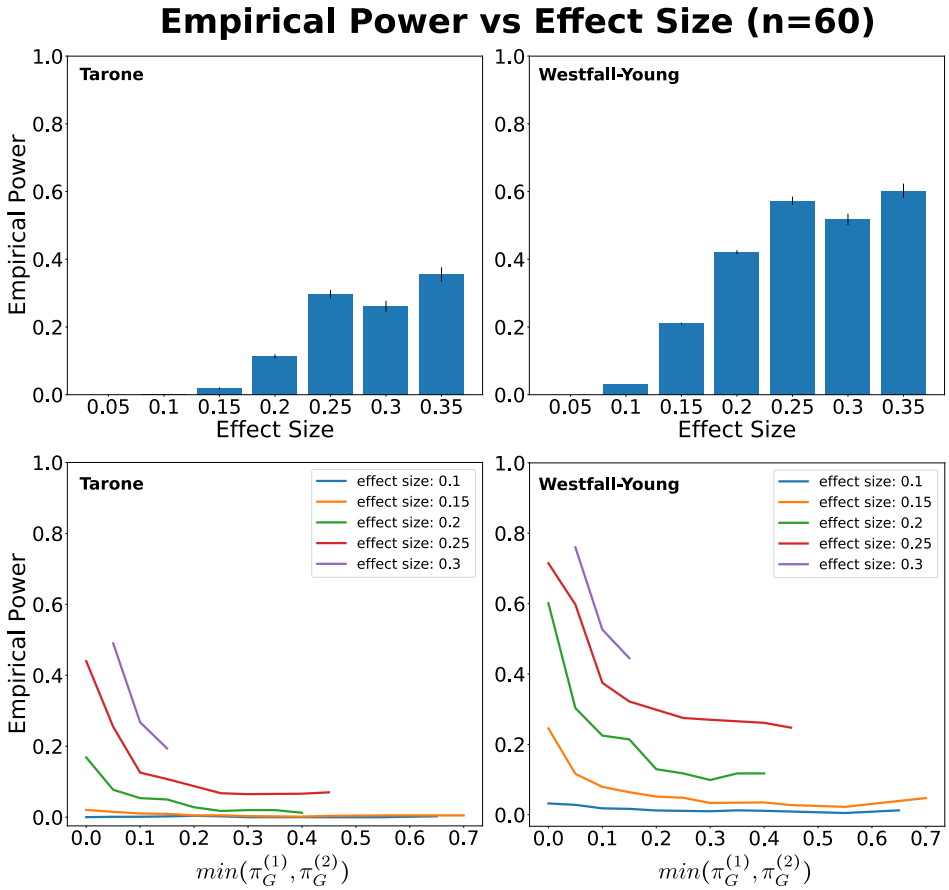

Fig. 6.12: Upper plots: Average empirical detection probabilities for subgraphs with different effect sizes (i.e. the average is over all subgraphs with a certain effect size and for each particular graph the detection probability is estimated as the fraction of detection among the 400 simulations). Error bars are plus minus one standard error. Standard errors were not calculated for effect size 0.05 due to computational constraints. There are more than 3.7 million subgraphs with this effect size meaning that in the order of 10^{12} detection covariances would have to be computed. This is necessary because the detections of different subgraphs are not independent. However, due to this large number of subgraphs, the standard errors are bound to be exceedingly small in this case. Lower plots: dependence of average detection probability on minimum of the two subgraph occurrence probabilities for different effect sizes. Even subgraphs with the same effect size have considerably different detection probabilities depending on how extreme the absolute occurrence probabilities are.

6.6 Discussion

What are the appropriate application cases for subgraph mining? A key feature of significant subgraph mining that distinguishes it from other statistical methods for graph comparisons is that it considers all possible differences between graph generating processes. In other words, as soon as these processes differ in any way, subgraph mining is guaranteed to detect these differences if the sample size is large enough. This is in contrast to methods that only consider particular *summary statistics* of the graph generating processes such as the average degree of a node. Such methods are of course warranted if there is already a hypothesis about a specific summary statistic. For example, [148] were specifically interested in the entropy of the distribution of shortest paths from a given node to a randomly chosen second node. In such a case, performing a statistical test with respect to the statistic of interest is preferable over subgraph mining because the multiple comparisons problem is avoided. This leads to a higher statistical power *regarding the statistic in question*. On the other hand, the test will have a low power to detect other differences between the processes. There are also well known methods such as the network-based statistic (NBS) developed by [149] operating on a more fine-grained level than summary statistic approaches. NBS aims to identify significant differences with respect to certain "connected components" of links. Thus, in terms of localizing resolution it is in between a summary statistic analysis and a full link-by-link comparison. Again, there is a trade-off here between statistical power with respect to certain features of the graph generating processes on the one hand, and resolution on the other. Compared to a method specifically tailored towards a particular summary statistic, the NBS will likely be less powerful. But due to its higher localizing resolution it will be able to detect differences towards which the summary statistic is blind.

Subgraph mining is on the far end of localizing resolution when it comes to comparing binary graph generating processes (by contrast NBS works with weighted graphs). Even if the two processes generate any individual link with the same probability there may be differences in terms of dependencies or interactions between link occurrences. These will be reflected in different subgraph probabilities for more complex subgraphs and subgraph mining is guaranteed to detect these differences given a sufficiently large sample. Of course, this comes at the price of having to deal with a very severe multiple comparisons problem. However, it would not be correct to say that for this reason subgraph mining has lower statistical power than more coarse-grained alternatives. Rather one should say that increasing the localizing resolution will always come at the price of a lower statistical power *with respect*

to certain differences but at the same time it will increase statistical power with respect to those differences that are only visible at the higher resolution. Given it's extremely high resolution we propose that subgraph mining should be the method of choice if no hypothesis about some specific difference between the graph generating processes is available so that no custom-tailored tests of those difference can be applied. In such a case subgraph mining can be utilized to systematically explore the entire search space of all possible differences.

What are the requirements on sample size? The appropriate sample size depends primarily on the kinds of effect sizes one seeks to be able to detect. Our empirical power analysis of the MEG data set discussed in the previous section suggests that in similar studies a sample size of about 60 is sufficient to have a very high probability to detect at least some of the existing differences. We carried out an additional analysis in order to narrow down the effect sizes likely to be detected at this sample size. This analysis showed that the largest effect sizes occurring in the empirical joint distribution (≈ 0.35 difference in probability of occurrence) had a detection probability of ≈ 0.4 on average using the Tarone correction and ≈ 0.6 on average using the Westfall-Young correction. This means that for a *particular* graph with a certain effect size the probability of detecting it is not extremely high. However, since there is generally a large number of such graphs there is a high probability of detecting at least some of them. Our analysis also showed that the effect size, understood as the difference in probability of occurrence of a subgraph between the groups, is not the only factor determining statistical power. Even graphs with the same effect size can have different probabilities of detection depending on how extreme the absolute probabilities of occurrence are. The detection probability is particularly high if the occurrence probability of a subgraph is close to zero in one of the groups. By symmetry we also expect this to be the case if it is close to one.

A possible way to reduce the amount of data required is to restrict the subgraph mining to subgraphs *up to a prespecified complexity* . For example, one could perform subgraph mining for all possible subgraphs consisting of up to three links. The validity of the method is not affected by this restriction. However, the search space is reduced and hence the multiple comparisons problem becomes less severe. In applying subgraph mining in this way it is important to pre-specify the desired complexity. Otherwise, we would run into yet another multiple comparisons problem. Consider the MEG data set presented in the previous section. Upon not detecting any differences with the full subgraph mining algorithm which considers all subgraphs on the seven nodes in our networks, one could check for differences among subgraphs consisting of at most six nodes. If nothing is found here either, we could move on to

five nodes and so forth until we are down to a single link comparison. However, this approach would not be valid because the individual links are essentially given seven chances to become significant so that our bounds on the family-wise error rate do not hold anymore.

What are the computational costs of subgraph mining? Besides the required sample size another factor for the applicability of subgraph mining is the computation time. The number of possible subgraphs can very easily be large enough that it becomes impossible to carry out a test for each one of them. Of course, the main idea behind the multiple comparisons methods presented here is that a large number of subgraphs can be ignored because they do not occur often enough or too often to be testable. For how many subgraphs this is true depends in particular on the connection density of the graphs. Generally, the computational load be will greater, the more nodes the graphs consist of and the more densely these nodes are connected. However, if the graphs are extremely densely connected one could revert to the *negative* versions of the graphs which would in this case be very loosely connected.

We provide a python implementation of significant subgraph mining as part of the IDTxl toolbox `http://github.com/pwollstadt/IDTxl` [103]. It offers both Tarone (with or without Hommel improvement) and Westfall-Young corrections. The latter is implemented utilizing the "Westfall-Young light" algorithm developed by [38] which we also adapted for within-subject designs. Details on the computational complexity can be found in this reference as well. The algorithm performs computations across permutations and achieves substantially better runtimes than a naive permutation-by-permutation approach. Our implementation is usable for both between-subjects and within-subject designs and allows the user to specify the desired complexity of graphs up to which subgraph mining is to be performed (see previous paragraph). It is also able to take into account the temporal network structure as described in the application to transfer entropy networks.

Which multiple comparisons correction method should be used? The choice between the two multiple comparison correction methods is a matter of statistical power on the one hand and a matter of false-positive control guarantees on the other. Regarding power, the Westfall-Young correction clearly outperforms the Tarone correction. Regarding false-positive control the situation is more complicated: whereas the Tarone correction is proven to control the family-wise error rate in the strong sense, the Westfall-Young procedure *in general* only provides weak control (see

[142]). There is, however, a sufficent (but not necessary) condition for strong control of the Westfall-Young procedure called *subset pivotality*. Formally, a vector of p-values $\mathbf{P} = (P_1, P_2, ..., P_m)$ has subset pivotality if and only if for any subset of indices $K \subseteq \{1, 2, ..., m\}$ the joint distribution of the subvector $P_K = \{P_i | i \in K\}$ is the same under the restrictions $\bigcap_{i \in K} H_{0i}$ and $\bigcap_{i \in \{1,...,m\}} H_{0i}$ [142, 150]. In the subgraph mining context this means that the joint distribution of p-values corresponding to subgraphs for which the null-hypothesis is in fact true remains unchanged in the (possibly counterfactual) scenario that the null-hypothesis is true for *all* subgraphs. It has been stated in the literature that the subset pivotality condition is not particularly restrictive and holds under quite minimal assumptions [151]. However, to the best of our knowledge, it has not yet been formally established in the subgraph mining context. A future study addressing this issue would therefore be highly desirable.

Just to clarify the practical role of the distinction between weak and strong control: weak-control does not allow a *localization* of differences between graph generating processes. It only warrants the conclusion that there must be *some* difference. The reason is essentially the same as the reason why it is not warranted to reject a null-hypotheses if it's p-value has not been corrected for multiple comparisons at all. Suppose we perform 20 tests at level 0.05 and a particular null hypothesis, say the fifth one, turns out to reach significance. If we did not correct for multiple comparisons, it would be a mistake to reject the fifth hypothesis because there is a plausible alternative explanation for why it reached significance: because we did not control for having performed twenty tests, it was to be expected that we would see at least one hypothesis rejected and it just *happened* to be the fifth one. Similarly, if we only have weak control of the FWER and a particular subgraph, say G_5, reaches significance, then it would be a mistake to conclude that G_5 is actually generated with different probabilities by the two processes. The alternative explanation is that our false positive probabilities are not controlled under the actual scenario (the ground truth) and G_5 simply happened to turn out significant. The only scenario that weak control *does* rule out (and this is how it differs from not controlling at all) is the one where all null-hypotheses are true, i.e. the one where the two graph generating processes are identical.

6.7 Conclusion

Significant subgraph mining is a useful method for neural network comparison especially if the goal is to explore the entire range of possible differences between graph generating processes. The theoretical capability to detect any existing stochastic difference is what distinguishes subgraph mining from other network comparison tools. Based on our empirical power analysis of transfer entropy networks reconstructed from an MEG data set we suggest to use a sample size of at least 60 subjects per group in similar studies. The demand on sample size and computational resources can be reduced by carrying out subgraph mining only up to a prespecified subgraph complexity or by reverting to the negative versions of the networks under consideration. The method can also be used for dependent graph generating processes arising in within-subject designs when the individual hypothesis tests and multiple comparisons correction methods are appropriately adapted. We provide a full python implementation as part of the IDTxl toolbox that includes these functionalities.

6.8 Supporting Information

6.8.1 Proof of validity of Tarone's correction factor

The validity of the Tarone correction factor $K(\alpha)$ can be seen as follows: Let $\mathcal{G}_0$ denote the set of subgraphs for which the null hypothesis of equal subgraph probabilities is true and let $T_0(k) = \{G \in \mathcal{G}_0 | p^*_G \leqslant \frac{\alpha}{k}\}$ be the subset of $\frac{\alpha}{k}$-testable subgraphs within $\mathcal{G}_0$. Furthermore, let $m_0(k)$ be the number of elements of this set, i.e. the number of $\frac{\alpha}{k}$-testable subgraphs for which the null-hypothesis is true. We can now compute the conditional family-wise error rate for a correction factor $k \in \mathbb{N}$ given the observed total frequencies of each subgraph. These frequencies can be interpreted as the realization of a random vector F containing one entry $F(G)$ per possible subgraph:

$$CFWER\left(\frac{\alpha}{k}\right) = \mathbb{P}\left(\bigcup_{G \in \mathcal{G}_0} \{p_G \leqslant \frac{\alpha}{k}\} \mid F = f\right) \tag{6.16}$$

We only have to take the union over $\frac{\alpha}{k}$-testable subgraphs because all other terms have probability zero:

$$= \mathbb{P}\left(\bigcup_{G \in T_0(k)} \{p_G \leqslant \frac{\alpha}{k}\} \mid F = f\right) \tag{6.17}$$

Using Boole's inequality (the "union bound"):

$$\leqslant \sum_{G \in T_0(k)} \mathbb{P}\left(p_G \leqslant \frac{\alpha}{k} \mid F = f\right) \tag{6.18}$$

By construction of the p-value we have for any constant $c \in \mathbb{R}^+ : \mathbb{P}\left(p_G \leqslant c | F = f\right) \leqslant c$. This fact can be applied to each term in the above sum with $c = \frac{\alpha}{k}$:

$$\leqslant \sum_{G \in T_0(k)} \frac{\alpha}{k} \tag{6.19}$$

The sum has $m_0(k)$ terms:

$$= m_0(k)\frac{\alpha}{k} \tag{6.20}$$

The number of testable subgraphs for which the null-hypothesis is true is smaller than or equal to the total number of testable subgraphs:

$$\leqslant m(k)\frac{\alpha}{k} \tag{6.21}$$

$$\overset{!}{\leqslant} \alpha \tag{6.22}$$

For the final equation (6) to be valid it must be true that $\frac{m(k)}{k} \leqslant 1$. In order to maximize the power of the resulting test, the correction factor should be chosen as small as possible. Hence, the appropriate choice is the smallest integer k such that $\frac{m(k)}{k} \leqslant 1$, i.e. $K(\alpha)$. Since the argument is valid for all possible observed total frequencies, it is also valid for the unconditional FWER which is simply a weighted average of the conditional FWERs:

$$FWER\left(\frac{\alpha}{K(\alpha)}\right) = \mathbb{P}\left(\bigcup_{G \in \mathcal{G}_0} \{p_G \leqslant \frac{\alpha}{K(\alpha)}\}\right) \tag{6.23}$$

$$= \sum_f \mathbb{P}(F = f)\mathbb{P}\left(\bigcup_{G \in \mathcal{G}_0} \{p_G \leqslant \frac{\alpha}{K(\alpha)}\} \mid F = f\right) \tag{6.24}$$

$$\leqslant \alpha \sum_f \mathbb{P}(F = f) \tag{6.25}$$

$$= \alpha \tag{6.26}$$

It is important to note that this argument does not make any assumptions about *which* or *how many* null-hypotheses are in fact true. The FWER is controlled in all cases. This property is called *strong control* of the FWER.

6.8.2 Hommel Improvement of Tarone's correction

The Tarone correction has been criticized on the basis that it is not α-consistent [152]. This means that a null-hypothesis might not be rejected at level α even

though it would have been rejected at an even smaller level $\delta < \alpha$. However, there is a simple modification proposed by [143] that makes the Tarone procedure α-consistent and, maybe more importantly, also improves its statistical power. The idea is to make the procedure α-consistent *by definition*, i.e. to reject H_0^G if the standard Tarone procedure would reject or if there exists a level $\gamma < \alpha$ such that the standard Tarone procedure would reject:

$$\text{Reject } H_0^G \text{ if and only if there exists a } \gamma,\ 0 < \gamma \leqslant \alpha, \text{ such that } p_G \leqslant \frac{\gamma}{K(\gamma)} \quad (6.27)$$

This rule has to be at least as powerful as the standard Tarone procedure because a null-hypothesis is rejected by the standard procedure it is also rejected by the improved version. Additionally, there are cases in which the Hommel impovement rejects but the standard Tarone procedure does not. Hommel presented a simple algorithm to implement this idea which, in the subgraph mining context, can be phrased as follows: First, we order all subgraphs in terms of their minimal achievable p-values such that $p^*_{G_1} \leqslant p^*_{G_2} \leqslant ... \leqslant p^*_{G_m}$, where $m = 2^{l^2}$ is the total number of possible subgraphs. Then we define the rejection rule as:

$$\text{Reject } H_0^G \text{ if and only if either } p_G \leqslant \frac{\alpha}{K(\alpha)} \text{ or } p_G < p^*_{G_{K(\alpha)}} \quad (6.28)$$

6.9 Acknowledgements

AG and MW are employed at the Göttingen-Campus Institute for Dynamics of Biological Networks (CIDBN) funded by the Volkswagen Stiftung. MW received support from the Volkswagenstiftung under the programme 'Big Data in den Lebenswissenschaften'. This work was supported by a funding from the Ministry for Science and Education of Lower Saxony and the Volkswagen Foundation through the "Niedersächsisches Vorab". MW received support from CRC 1193 C04 funded by the DFG. We thank Lionel Barnett for helpful discussions on the topic.

6.10 Author contributions

Aaron Julian Gutknecht: Conceptualization; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – original draft; Writing – review & editing. Michael Wibral: Conceptualization; Data curation; Project administration; Resources; Supervision; Writing – review & editing.

7 General discussion

In the previous five chapters we have ventured from the statistics of linear information flow, to the extension of information theory through Partial Information Decomposition, to the theory of graph comparisons allowing us to find differences in patterns of information flow. In the following section, we highlight the key insights derived from this thesis and outline some immediate directions for future research. Subsequently, in Section 7.2, we explore how the concept of information and the specific approaches discussed in this thesis, relate to other important concepts in the analysis of complex systems. The thesis concludes with some final remarks in Section 7.3.

7.1 Key insights and future directions

7.1.1 Granger causality

In Chapter 2, we introduced the previously unknown asymptotic sampling distributions for 'single regression' Granger causality estimators, both in the time domain and for spectral Granger causality averaged over a frequency band of interest. These estimators have a significant relevance in diverse fields such as econometrics and neuroscience addressing notable shortcomings of the standard log-likelihood ('dual regression') estimators. The chapter yields the following key insights:

Key insights Chapter 2

1. The asymptotic (large sample) sampling distribution of single regression Granger causality estimators is a generalized χ^2 distribution in both time and (band-limited) frequency domains. The generalized χ^2 distribution is well approximated by a Γ distribution.

2. Valid significance tests can be constructed based on these distributions by first projecting the estimated model parameters into the null-space and then using the $1-\alpha$ quantile of the corresponding null-distribution as a rejection cut-off.

3. The analytical method underpinning our main results, a multivariate second-order Delta method, offers remarkable versatility and can be used to derive sampling distributions in other important cases of interest within Granger causality analysis.

We have provided analytical expressions for the parameters of the sampling distributions in the case of unconditional Granger causality for finite-order vector auto-regressive (VAR) models. We also explored the error-statistical properties (type I and statistical power) in this setting. Overall, the time-domain test offers similar statistical power as a likelihood ratio test. Yet there are parameter regimes in which one or the other test may be preferable. This offers the possibility to develop improved "two-stage" tests where we first identify the parameter regime and then select the most appropriate test for that specific regime. However, this possibility has yet to be explored.

For technical reasons explained in the main chapter, the null-hypothesis of vanishing Granger causality over a frequency band is in fact identical to the time-domain null-hypothesis. However, the test based on the band-limited estimator offers insights into the distribution of Granger causality over the frequency spectrum that the time-domain test does not. This is due to a difference in the power profiles of the tests: whereas the time-domain test is primarily concerned with the overall magnitude of Granger causality, the band-limited test is selectively sensitive to the magnitude of Granger causality within the frequency-band of interest.

Under specified conditions, our analytical method yields a generalized χ^2 distribution in a range of other important cases as well. While the discussion section elaborates on various cases where our analytical method can be applied, two specific extensions merit particular emphasis for future research. First, the extension to conditional Granger causality, frequently used to evaluate pairwise Granger causality between

nodes in a network while conditioning on all other nodes. Second, the extension to general state-space models, which are equivalent to vector auto-regressive moving-average (VARMA) models. This extension is very desirable because many real-world data contain a strong moving-average component. In both of these cases, the asymptotic sampling distribution of the respective estimators will be a generalized χ^2 distribution. Its specific parameters as a function of the model parameters remains to be determined however.

7.1.2 Partial Information Decomposition

The work presented in Chapters 3, 4, and 5 jointly advance our understanding of Partial Information Decomposition, from a concrete way to measure the atoms of information to its abstract mathematical structure.

Chapter 3 introduces the shared-exclusions measure i^{sx} of pointwise redundant information, i.e. of the information redundantly carried by a set of specific source realizations $s_1, \ldots, s_n$ about a target realization t. This, in turn, automatically induces a full Partial Information Decomposition (see Section 3.4). Key insights gained in this Chapter are

Key insights Chapter 3

1. Redundant information can be measured in terms possibilities being excluded jointly by all source realizations, and equivalently, as the information provided about the target by certain logical statements about the source realizations.
2. The resulting measure has the form of a pointwise mutual information, inheriting its interpretation and mathematical properties.
3. It is differentiable with respect to the underlying source-target joint probability distribution and obeys a target chain rule making it particularly useful for practical applications.
4. The measure induces a full Partial Information Decomposition, i.e. a measure of each information atom. This induced PID is differentiable as well.

One immediate direction for future research regarding the i^{sx} measure lies in the development of statistical tests for the measure itself and its implied PID. Fundamentally, this is a problem in non-parametric statistics. We are dealing with a

functional of the joint probability distribution between source and target variables and in general we do not wish to impose any parametric form on that distribution. Methods such as the non-parametric bootstrap would be natural candidates for this purpose. Such methods have yet to be fully explored in this context, however. In many cases, we may only be interested in the relative contributions of different types of information, i.e. their percentages of the total mutual information. This question falls under the umbrella of "compositional data analysis", which is concerned with statistics for quantities that are subject to a constant sum constraint [153]. Exploring this line of research could provide valuable insights and methods for PID analysis.

The original i^{sx} measure, as introduced in Chapter 3, is designed for discrete variables. However, many applications in neuroscience and other fields are better modeled using continuous variables. Anticipating this, the measure-theoretic formulation provided in the appendix of the chapter laid the groundwork for extending i^{sx} to a continuous setting. In subsequent research, this continuous extension has been successfully achieved using a measure-theoretic approach [154]. Moreover, the extension is also able to handle mixed discrete-continuous data. Notably, the continuous version of i^{sx} is transformation invariant under invertible mappings, a property often deemed highly desirable for information measures. Despite these advancements, significant questions remain for future research, such as the analytical evaluation of this measure for concrete multivariate continuous probability distributions, as well as the development of efficient estimators.

Since its introduction the i^{sx} measure has found promising applications in the realm of artificial neural networks which we briefly review in the following. Firstly, the i^{sx} measure has been leveraged in a study by Ehrlich et al. [30] to investigate the 'representational complexity' of artificial neural networks. Recall from the introduction (Section 1.3.2) that representational complexity can be defined in terms of PID atoms and quantifies, roughly speaking, the average number of nodes in a layer required to decode the target, which in this context is the correct output label. Ehrlich et al. applied this measure to analyze the hidden layer representations of neural network classifiers tackling well-established tasks such as MNIST handwritten digit recognition and CIFAR10 image classification. Here, their key finding is that representational complexity decreases both over the course of training and through successive layers. The study demonstrates how representational complexity, derived from i^{sx}, can offer valuable insights into the structure of internal representations within neural networks which go beyond simple pairwise measures such as mutual information.

Secondly, the i^{sx} measure has been utilized in the context of 'infomorphic networks,' which aim to optimize abstract information processing goals framed within the PID framework (as discussed in Section 1.3.2). The differentiability of i^{sx} makes it the ideal measure for such applications, enabling the analytical calculation of the gradients required for training. Graetz et al. [34] showed how i^{sx}-derived goal functions can be successfully employed in various learning paradigms including supervised and unsupervised settings, as well as associative memory learning. In the latter context, the network strives to memorize specific patterns and aims to reconstruct them when presented with distorted or incomplete versions.

Let us move on to Chapters 4 and 5 concerned with the mathematical structure of PID. Chapter 4 presents a derivation of PID theory from first principles. It seeks to develop PID theory on the basis of two of the most elementary relationships of human thought: the part-whole relationship and the relation of logical implication. In contrast to the original formulation, this approach addresses the problem directly from the perspective of the information atoms, i.e. the quantities we are ultimately interested in, instead of introducing these quantities indirectly through cumulative measures such as redundant information. Firstly, this has the advantage of making the intended interpretation of the information atoms in the general n-sources case particularly clear. Secondly, it illuminates the logical structure of the PID problem and the theoretical roles played by concepts such as antichain lattices. Thirdly, it makes certain important questions in PID theory very easy to address, in particular the question of the construction of non-redundancy based PIDs. The key insights provided by this chapter are:

Key insights Chapter 4

1. Each information atom in a PID can be defined by its characteristic parthood relationships with respect the information provided by the different possible subsets of source variables about the target. This approach can be formalized via the concept of parthood distributions.
2. Equivalently, PID theory can be explained in terms of a hierarchy of logical constraints on how certain pieces of information about the target can be accessed. This approach can be formalized via the concept of statements with monotonic truth-tables and proves useful for deriving properties of PID lattices.
3. The axioms of the original formulation can be proven as theorems within the novel approach.
4. The shared exclusions measure of redundancy i^{sx} can be re-derived using principles of logic and mereology (the study of part-whole relations).
5. The novel approach can be utilized to systematically address the question of non-redundancy based PIDs, such as synergy-based PIDs.

The PID problem is, in essence, a problem of *explication*, a term coined by Rudolf Carnap in his seminal book "Logical Foundations of Probability" [155]. It is the problem of translating some vaguely defined concepts of some earlier stage of scientific development ("uniqueness", "redundancy", "synergy"), into an exact and, in this case, formal-mathematical language. One interesting aspect of the PID problem that distinguishes it from other standard problems of explication (such as the explication of the notion of probability) is that it is concerned with the simultaneous explication of a large number of quantities at the same time. Moreover, these quantities are characterized not only by their intrinsic properties (e.g. properties we might expect a measure of "redundancy" to satisfy) but, crucially, also by their *relationships* to each other and to previously introduced quantities, in particular mutual information. The approaches presented in Chapter 4 emphasize this relational aspect of the PID problem and aim to formalize the relevant relationships ("What is contained in what?") using notions such as "parthood distributions" f, "parthood lattices" $\mathcal{B}$ and "parthood conditions" $\mathcal{C}(\alpha, f)$.

Chapter 5 directly builds upon these insights, utilizing the parthood approach to construct a general scheme of *PID base-concepts*. These are information functionals

that imply a full PID once they are defined in terms of the underlying source-target joint distribution (such as redundant information):

Key insights Chapter 5

1. Within the parthood approach base-concepts can be expressed in terms of conditions phrased in formal logic on the specific parthood relations between the PID atoms and the different mutual information terms.
2. There is a general pattern of such logical conditions that encompasses all base-concepts discussed in the literature and also leads to novel base concepts.
3. The pattern categorizes base-concepts into four equivalence classes of "partner measures" quantifying the same information components but viewed from the perspective of different source collections. The equivalence classes can be represented by the concepts of 1) redundant information, 2) weak synergy, 3) union information, and 4) vulnerable information. Additionally, inducing a PID using unique information as a base measure is possible, but amounts to defining the information atoms directly.
4. The concept of vulnerable information quantifies information that may be lost if we lose access to one of the sources. It has not been considered in the literature before.

Recall from the main chapter (Section 5.1) that the question of PID base-concepts serves *interpretational*, *computational*, and *theoretical* purposes. Firstly, if a PID is constructed using a specific base-concept, we gain immediate control over the interpretation of that base-concept, as it is directly defined in terms of the joint distribution. Secondly, the choice of base-concept can have computational advantages. For example, if we are mainly interested in quantities related to the idea of synergistic information, using synergy as a base-concept is generally computationally advantageous. A case in point is the measure of representational complexity we discussed above and in Section 1.3.2. The computational cost of computing this measure scales as $\mathcal{O}(n)$ (n being the number of source variables) when using synergy as a base-measure. Using redundancy, on the other hand, would necessitate computing all PID atoms, resulting in super-exponential scaling. Thirdly, understanding the relationships between different base-concepts can provide important theoretical insights. Because base-concepts are in an invertible relation with the information atoms—and consequently in a one-to-one relation with each other—imposing axioms on one

base-concept will inevitable imply certain properties for the other base-concepts (or imply that they cannot have certain properties). This type of exploration could be useful for narrowing the set of possible solutions to the PID problem.

One starting point in this context could be a thorough exploration of the novel base-concept of vulnerable information. As a measure of the "non-robustness" of representations, understanding vulnerable information could have significant implications for systems where resilience is critical. For instance, if most of the information in a neural network about a specific aspect of the external world is vulnerable, then disruptions to the system could lead to immediate loss of that information. The next step in this line of research is to formally define a measure of vulnerable information in terms of the underlying probability distribution. Key considerations would include determining what reasonable axioms such a measure should satisfy to adequately capture its intuitive meaning.

Another interesting avenue to explore is whether the constructions of Chapters 4 and 5 can also be used to address similar decomposition problems. The information relationship ("sources providing information about a target") might not be the only relationship where a decomposition into unique, redundant, and synergistic components (and their appropriate generalizations) can be achieved. Consider for example logical-deductive relationships between statements, i.e. premises $A_1, \ldots, A_n$ implying some conclusion C. Here one could similarly distinguish between unique, redundant, and synergistic *inferences*, i.e. inferences to conclusions that uniquely follow from particular premises, that redundantly follow from multiple premises, or that follow only synergistically utilizing multiple premises at the same time. The number of possible types of inferences from n premises should be the same as the number of information atoms in a PID. The reason is that just like "information", the relation of "deductive inference" is monotonic: everything that follows from a subset of premises will also follow from any superset of it. Accordingly, there will be one type of inference per (non-constant) monotonic Boolean function $f : \mathcal{P}(\{A_1, \ldots, A_n\}) \to \{0, 1\}$ describing which subsets of premises the conclusions of the type described by f follow from (similar to PID, the constant Boolean functions only lead to trivial cases).

A similar decomposition problem arises in the realm of causal relationships where one might distinguish between unique, redundant, and synergistic *effects*. This is discussed below in Section 7.2.2. Let us now turn the final Chapter of this thesis about subgraph mining as a statistical tool for network comparison.

7.1.3 Subgraph mining

Chapter 6 addresses the issue of statistically comparing two stochastic processes that generate unweighted graphs defined on the same set of nodes. These processes may stand for two separate experimental groups, each producing one graph per subject in the respective group. The graph nodes could represent specific brain locations where activity levels are measured through a given neuroimaging technique, and edges might indicate the presence of statistically significant Transfer Entropy between these nodes.

The primary objective is to discern stochastic differences between the two processes, i.e. differences concerning the probabilities with which they generate certain edge patterns, or *subgraphs*. Individual edges are also included in this analysis. However, the differences between the processes can also rest in more complex patterns, reflecting differences in the dependencies between edges. Given that even a moderate number of network nodes can result in an immense variety of such patterns, a severe multiple comparisons issue arises. The method of Significant Subgraph Mining was recently introduced [38, 39] to address this problem for independent graph-generating processes, offering guarantees on false positive rates while examining the entire space of possible patterns. Chapter 6 adapts and extends this method, specifically for the comparison of functional connectivity networks in neuroscience. The key insights of the chapter are as follows

Key insights Chapter 6

1. Subgraph mining serves as a valuable follow-up analysis after network inference methods like Transfer Entropy, or Granger causality have been applied across multiple groups or conditions. In contrast to other network comparison tools, it is capable of detecting any existent differences between graph-generating processes given a sufficiently large sample size.

2. The method can be extended to handle dependent graph-generating processes, making it applicable in within-subject experimental designs.

3. Based on empirical power analysis of Transfer Entropy networks, a target sample size of approximately $n = 60$ is advised for a high probability of detecting at least some existing differences in similar studies.

4. An open-source Python implementation is made available through the IDTxl toolbox, which is tailored for neuroscience research. This includes features like the inclusion of information transfer delays in the graph structures and the ability to limit subgraph complexity, thereby reducing computational costs.

The chapter discussed two different methods for multiple comparisons correction: the Tarone correction and the Westfall-Young correction. While the Westfall-Young correction appears more powerful in simulations, it's not yet clear whether this advantage comes at the cost of weaker error-control guarantees. Specifically, the Tarone correction ensures strong control of the family-wise error rate—control irrespective of which and how many null hypotheses are true. In contrast, the Westfall-Young correction is known to offer weak control – control under the condition that all null hypotheses are true. A technical condition known as 'subset pivotality' would provide strong control for the Westfall-Young method as well, but it remains an open question whether this condition holds in the case of subgraph mining. Clarifying this question would broaden the method's applicability, especially for studies with smaller sample sizes.

7.2 Beyond information: knowledge, causality, and utility

7.2.1 Information and knowledge

The shared exclusions measure introduced in Chapter 3 and rederived in Chapter 4 has tight connections to epistemic logic, i.e. the logic of reasoning about knowledge [156]. This becomes particularly apparent in the "event-based approach" to multi-agent epistemic logic described in [157]. The basic mathematical structure in this approach is called an *Aumann structure*. It consists of a set Ω of states of the world and a set of partitions $\mathcal{P}_1, \ldots, \mathcal{P}_n$ of Ω, one for each of n agents. We can imagine each partition arising from the agents making observations described by an observation function $S_i : \Omega \to \mathcal{S}_i$ where $\mathcal{S}_i$ is the alphabet of possible observations agent i can make [79]. Each state of the world ω gives rise to a specific observation s_i for agent i. From this observation agent i can infer that the actual state of the world must be in the subset of states giving rise to the specific observation s_i (i.e. the state must be in the preimage $S_i^{-1}(s_i)$ of s_i under S_i). In this way, for each agent, the states of the world Ω are partitioned into equivalence classes of states indistinguishable by the agents based on their observations:

$$\mathcal{P}_i = \{S_i^{-1}(s_i) : s_i \in \mathcal{S}_i\} \tag{7.1}$$

For any event $E \subseteq \Omega$ we can say that agent i *knows* E just in case the event E occurs in all the states that appear possible to the agent given their observation s_i:

$$S_i^{-1}(s_i) \subseteq E \text{ ("agent i knows E")} \tag{7.2}$$

We can now think about the set of states of the world such that *all* agents know that E has occurred, i.e. the set of states such that it is *shared knowledge* that E has happened. Given the terminology just introduced this can be defined as

$$SK(E) := \{\omega \in \Omega : \forall i \text{ agent i knows E given observation } S_i(\omega)\} \tag{7.3}$$

The connection to the shared exclusions measure of redundant information can be drawn by equipping the set of states Ω with a probability measure $\mathbb{P}$ and $\sigma-$algebra $\mathfrak{A}$ (or in other words, interpreting Ω as a sample space in a probability model). The observation functions S_i become random variables in this case. Let us introduce an additional random variable $T : \Omega \to \mathcal{T}$ representing some aspect of the state of the

world. For instance, T could describe the weather in a given state of the world and take on values "sunny", "rainy", "foggy", etc.

Now we may ask: how much information does the entirety of the agents' shared knowledge about the state of the world provide about T? The entirety of shared knowledge can be captured via the *union* of the preimages of the agent's observations: Given observations $s_1, \ldots, s_n$, the agents have shared knowledge about the union of the events $S_i^{-1}(s_i)$ since all agents can infer, given their observations, that the actual state of the world must lie in this union. Further, this is the strongest, and hence most informative, restriction of the states of the world that all agents can agree upon. Thus, it makes sense to say that the agents' shared knowledge about the event $E^* = \bigcup S_i^{-1}(s_i)$ is in fact all of their shared knowledge about the state of the world.

For this reason it seems plausible to compute the information about the actual value t of T provided by the agents' shared knowledge about the state of the world as the pointwise mutual information

$$i(\mathbf{1}_{E*} = 1 : t) = \log\left(\frac{\mathbb{P}(T = t|\bigcup S_i^{-1}(s_i))}{\mathbb{P}(T = t)}\right) \tag{7.4}$$

where $\mathbf{1}_{E*}$ is the indicator of event E^*. This measure essentially says: if we relied exactly on what is known by all agents, namely that the actual state of the world is in $\bigcup S_i^{-1}(s_i)$, by what factor would that make it more or less likely to guess the correct target value t (compared to not taking into account the agents' observations at all)? However, this is precisely $i^{sx}(s_1, \ldots, s_n : t)$. Hence, i^{sx} quantifies *the information about the target realization provided by the shared knowledge of the agents about the state of the world*.

7.2.2 Information and (interventional) causality

Being firmly rooted in probability theory, information theoretic analyses, including Transfer entropy / Granger Causality and PID, are concerned with *stochastic* dependencies and the underlying notion of Wiener-Granger causality is one of purely *predictive* causality as explained in Section 1.2. This leaves open what the underlying *interventional causal* structure of the system might look, i.e. how does the system behave under interventions on some of its components? The theory of interventional causality has been formalized most prominently in terms of Structural Causal Models

(SCMs) [158]. Given variables of interest $X_1, \ldots, X_n$, an SCM $\mathfrak{E}$ consist of a set of structural assignments

$$X_i := f_i(\mathbf{PA}_i, U_i) \tag{7.5}$$

where $\mathbf{PA}_i \subseteq \{X_1, \ldots, X_n\} \backslash \{X_i\}$ are called the parents of X_i and the variables U_i are mutually independent noise variables with a given joint distribution $P_{U_1,\ldots,U_n}$ [158, 159]. Any SCM is associated with a causal graph, where the variables X_i are the nodes and each X_i has incoming arrows from all its parent nodes. If the causal graph is acyclic (which is by far the most intensively studied case in the literature), the distribution over the noise variables automatically entails a joint distribution $P^{\mathfrak{E}}_{X_1,\ldots,X_n}$ over the X_i via the structural assignments. This is usually called the *observational distribution*.

Interventions are modelled in this framework as replacements of certain structural equations. For instance, one could intervene on the system by setting X_1 to a specific value, say, zero. In this case, the equation for X_1 would be replaced by "$X_1 := 0$" leading to a new structural equation model denoted in terms of the "do"-operation as $\mathfrak{E}; do(X_1 := 0)$. This notation emphasizes that we are actively changing the underlying system. The new entailed joint distribution over the X_i is called the *interventional distribution* $P^{\mathfrak{E};do(X_1:=0)}_{X_1,\ldots,X_n}$. There is said to be a *causal effect* from X_i to X_j just in case there is a value x_i such that intervening on X_i by setting it to x_i changes the entailed distribution of X_j, i.e. if $P^{\mathfrak{E}}_Y \neq P^{\mathfrak{E};do(X_1:=0)}_Y$ [159].

Under certain conditions, this interventional conception of causality aligns with the Wiener-Granger conception framed in terms of conditional independence (or equivalently, vanishing Transfer Entropy). Specifically, Peters et al. [159] show that if we have an SCM for a stochastic process $\mathbf{X}_t$ ($t \in \mathbb{Z}$) with *no instantaneous effects*, and the entailed distribution is *faithful* with respect to underlying causal graph, then there will be a causal arrow from one variable to another in the causal *summary graph* if and only if there is non-zero Transfer Entropy from the former variable to the latter. In this sense Wiener-Granger causality is the right condition for interventional causality given the above assumptions.

Let us unpack this statement. In contrast to the SCMs discussed so far, an SCM for a stochastic process is defined over an infinite number of variables and accordingly the *full causal graph* of the model will contain infinitely many nodes. Since causal relationships should not go from future to past, the causal graph should only contain an arrow from $X_{i,t}$ to $X_{j,t'}$ if $t \leqslant t'$. If there is a causal arrow between variables at the same time index, this is called an *instantaneous* effect. The above statement excludes this case. The causal structure can be summarized into a *summary graph* with only n nodes, one for each component process. This summary graph will have

an arrow from X_i to X_j just in case there is a causal arrow $X_{i,t}$ to $X_{j,t'}$ for *some* $t \leqslant t'$, i.e. just in case there is some causal influence of the i-th component process on the j-th component process.

The *faithfulness* assumption is central to many state-of-the-art causal inference techniques. It relates properties of the causal graph to conditional independence properties of the observed joint distribution. Specifically, faithfulness of an SCM means that if two sets of variables are conditionally independent given a third set of variables (all three sets are disjoint), then in the causal graph the first two sets are *d-separated* given the third set. d-separation is a purely graph theoretic concept defined in terms of the conditions under which paths between nodes of the first two sets are "blocked" by nodes in the third set. The concept is formulated in such a way that for the causal graph of an SCM, d-separation automatically implies conditional independence in the observational distribution. Faithfulness is the other direction of this implication allowing an inference from observed conditional independencies to the equivalence class of causal graphs satisfying the corresponding d-separation statements.

Intriguing connections also exist between interventional causality and the Partial Information Decomposition problem. Much like the concepts of 'unique,' 'redundant,' and 'synergistic' information had been circulating in scientific discourse prior to their formal treatment by Williams and Beer [19], terms such as 'unique causation' and 'synergistic causation' are also quite prevalent (a quick search for the phrase 'synergistically cause' yields over a thousand results on Google Scholar). This naturally raises the question: Is it possible to achieve a similar formalization for these types of causation? Moreover, could insights garnered from PID theory help improve our understanding of these distinct forms of causality?

Two distinct projects can be distinguished in this context. Firstly, we may ask how to decompose some measure of the total causal effect of causal factors $C_1, \ldots, C_n$ on some target variable T into unique, redundant, and synergistic components (and appropriate generalizations thereof)? This may be called a *Partial Causation Decomposition (PCD)*. Chapter 4 formulated three fundamental questions that have to be addressed in order to systematically construct a PID. These can be translated into the causal domain as follows

1. What do the components of a Partial Causation Decomposition mean, i.e. what is their intended interpretation?
2. How many components are there for a given number of causal factors?
3. How to quantify the different components?

The first two questions pertain to the structure of a PCD. This structure will likely be different from PID because there is a key difference between causal and informational relations: while informational relations are monotonic (by adding an information source we can only gain information), causal relations are non-monotonic. Adding a causal factor may prevent certain effects that might otherwise have happened. To illustrate this, imagine I throw a stone at a window. However, you simultaneously throw another stone that intercepts mine mid-air, preventing the window damage that would have occurred otherwise. The answer to the third question will heavily depend on how exactly we measure the "total causal effect", i.e. the *strength* of causal relationships. Unlike PID where we have a generally accepted measure of the quantity we wish to decompose, namely mutual information, the same is not the case in the causal domain (but see in particular [160]).

A second project relating PID and interventional causality would be whether the different information atoms in a PID can be given a causal interpretation, i.e. to what extent the atoms *themselves* are causally relevant with respect to the target variable. For instance, if we intervened in such a way that, say, the synergistic information provided by the source variables about the target vanishes, how would that affect the distribution of the target? Can we even measure *in bits* how much of the information in each atom is causally relevant to the target variable? This might be called *Causal Information Decomposition*. The key question in this regard would be how to construct an appropriate class of interventions affecting specifically certain information atoms otherwise leaving everything unchanged.

7.2.3 Information and functional utility

In essence, information theoretic quantities capture stochastic dependencies between various components of a system (e.g., neurons in the brain) or between a system and its external environment (e.g., a neural network and an environmental variable it encodes). These dependencies enable an external observer to make predictions about the system's behaviour or about the environment based on observations of the system. However, especially in the context of biological systems such as the brain, one must ask how the system *itself* utilizes this information to achieve specific tasks or goals.

Previous research, notably by Bialek [161], has explored the relationship between information and function. This work demonstrates that achieving a specific level of performance—which can be thought of as the system's ability to achieve a given

goal—requires the system to have access to a certain minimum amount of information. Bialek uses the illustrative example of bacteria in an environment where the availability of sugar can vary. To metabolize the sugar optimally, the bacteria should produce an enzyme in a quantity finely tuned to the available sugar, ensuring there is just enough for metabolism without exceeding this amount, as excess production would consume valuable energy. In this example, the amount of available sugar can be considered as an 'external state' (Y), and the amount of enzyme produced by the bacteria can be viewed as an 'internal state' (X). The growth rate λ, which we can think of as a measure of performance, can then be conceptualized as a function of both these internal and external states ($\lambda = \lambda(X, Y)$).

It can now be formally shown that a given level of performance $\lambda(X, Y) = \lambda_0$ is attainable only if the bacterium has access to a minimal amount of information $I_0(\lambda_0)$ about its environment, i.e., if $I(X : Y) > I_0(\lambda_0)$. Importantly, a system may possess more information about its environment than is strictly necessary for its current performance level. In such cases, it appears that the system is not fully utilizing the available information, or possibly, it may have the wrong information in some sense. This leads to the realization that not all information is functional in terms of improving performance, suggesting that there may be non-functional or 'excess' information as well.

Given the considerations laid out thus far, a pressing challenge emerges: to bridge the gap between purely predictive informational relationships and the functional utilization of information by the system itself. While information-theoretic metrics like Granger Causality, Transfer Entropy, and Partial Information Decomposition (PID) provide valuable insights into predictability, they don't —in and of themselves— speak to functionality. Thus, an important interpretational question arises: to what extent is the measured information actually being functionally utilized by the system under study and how can we quantify this extent?

There have been some approaches to this question in the literature. For example Polani et al. [162] propose a measure of the "relevant information" in an environmental variable Y. This could be thought of as an upper bound on the functional information we have been discussing so far. "Information that may be used", even if it in fact isn't. Roughly, the measure can be understood as follows: imagine that the system can perform actions the utility of which depends on the state Y. Since the agent has some uncertainty about the environment, the agent will thus also have some uncertainty about what the optional action would be in the case at hand. The relevant information in Y describes to what resolution the system needs to know the value of Y in order to perform the optimal action. In the extreme case, the system

might need to know the exact value of Y in order to choose the right action. In other words, the system needs to know Y to a resolution of $H(Y)$ bits. However, suppose that for the choice of the right actions the system only needs to know whether the value of Y falls into one of two equally likely partitions of its possible values. In that case, the system only needs to know Y to a resolution of 1 bit, and accordingly, this would be the amount of relevant information in Y in that case.

There have also been approaches directly concerned with the notion of functional information discussed above. In particular, Kolchinsky [163] proposes a measure of "semantic information" that a system has about its environment, which is understood as "information which is in some sense meaningful for a system". Kolchinsky's approach fits nicely with the above discussion about performance functions as well as our discussion regarding interventional causality in Section 7.2.2. The intuition here is that semantic information has be to causally relevant to the operation of the system, i.e. it must "make a difference" in the sense that intervening on the informational relationship is disadvantageous to the system. To what extent this is the case can be formally expressed in terms of a "viability function". We are then intervening on the system-environment relationship so that the information the system carries about the environment is scrambled. The question is: how much of this information can we scramble away so that the viability remains unchanged? Kolchinsky defines the semantic information about the environment as the remaining information under the "viability optimal intervention", i.e. the intervention which scrambles away the most information without affecting the viability.

These consideration open very interesting possibilities for future research in particular in relation to PID theory. Firstly, since the measure of relevance by Polani et al. is phrased as a mutual information between an environmental variable and a "relevance indicator variable", representing the action chosen by an optimal agent, PID theory could be directly applied here if the environmental variable has multiple components $Y_1, \ldots, Y_n$. These can be thought of as source variables. Taking the relevance indicator variable as the target we obtain a *decomposition of relevance* telling us where the relevance in the complex environmental variable resides. To what degree is it in a particular component? To what degree is it redundantly contained in multiple components? And to what degree does it reside synergistically in some collective property of the environmental variable?

Secondly, we may ask if something like a *functional information decomposition* is possible. Given a decomposition of the information a system carries about its environment (or perhaps about other parts of the system), is it possible to explicitly measure the extent to which the different information atoms functionally contribute

to the systems' performance? The approach by Kolchisnky suggest to obtain such a decomposition in terms of PIDs associated with viability optimal interventions. This would certainly be a fruitful avenue to transition from a predictive to a functional understanding of information.

7.3 Concluding remarks

This thesis has tackled a variety of current topics in information theory, particularly as they relate to the analysis of complex networks. We began by presenting new results in the statistical theory of linear information flow, deriving the asymptotic null-distribution for single regression Granger causality estimators.

Our exploration then moved to Partial Information Decomposition (PID) theory, a significant extension of classical information theory with numerous promising applications. Utilizing insights from formal logic and mereology—the study of part-whole relationships—we derived PID theory from first principles and introduced a general logical scheme of PID base-concepts, i.e. PID-inducing information functionals. Furthermore, we offered a concrete solution to the PID problem in the form of the shared exclusions measure of redundant information i^{sx}.

Lastly, we addressed the issue of network comparison, which often emerges as a subsequent step in information-theoretic analyses. We demonstrated how to adapt and extend the method of Significant Subgraph Mining, particularly for applications in information-theoretic network inference in neuroscience.

In summary, all of these topics contribute important elements towards a comprehensive theory of multivariate dependencies, both in biological and non-biological networks. They open up a wide range of promising research directions and establish connections to other key areas such as epistemic logic, interventional causality, and the exploration of functional information. The integration of these diverse yet interconnected topics highlights the richness and complexity of the field, suggesting an exciting path forward.

Bibliography

[1] Claude E Shannon. “A mathematical theory of communication”. In: *The Bell system technical journal* 27.3 (1948), pp. 379–423 (cit. on pp. 2, 108).

[2] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999 (cit. on pp. 3, 108).

[3] William McGill. “Multivariate information transmission”. In: *Transactions of the IRE Professional Group on Information Theory* 4.4 (1954), pp. 93–111 (cit. on pp. 6, 106).

[4] Joseph T Lizier, Mikhail Prokopenko, and Albert Y Zomaya. “Detecting non-trivial computation in complex dynamics”. In: *European Conference on Artificial Life*. Springer. 2007, pp. 895–904 (cit. on p. 7).

[5] Terry Bossomaier, Lionel Barnett, Michael Harré, et al. *Transfer entropy*. Springer, 2016 (cit. on pp. 7, 8).

[6] T. Schreiber. “Measuring information transfer”. In: *Phys. Rev. Lett.* 85.2 (2000), pp. 461–4 (cit. on pp. 7, 27).

[7] Norbert Wiener and Edwin Beckenbach. “Modern mathematics for engineers”. In: *New York: McGraw-Hill* (1956) (cit. on p. 8).

[8] Anil Seth. “Granger causality”. In: *Scholarpedia* 2.7 (2007), p. 1667 (cit. on p. 8).

[9] J. Geweke. “Temporal aggregation in the multiple regression model”. In: *Econometrica* 46.3 (1978), pp. 643–661 (cit. on p. 8).

[10] J. Geweke. “Measurement of linear dependence and feedback between multiple time series”. In: *J. Am. Stat. Assoc.* 77.378 (1982), pp. 304–313 (cit. on pp. 8, 24–28, 32, 45).

[11] J. Geweke. “Measures of Conditional Linear Dependence and Feedback Between Time Series”. In: *J. Am. Stat. Assoc.* 79.388 (1984), pp. 907–915 (cit. on pp. 8, 24, 28, 45).

[12] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag, 2005 (cit. on p. 8).

[13] L. Barnett, A. B. Barrett, and A. K. Seth. “Granger causality and transfer entropy are equivalent for Gaussian variables”. In: *Phys. Rev. Lett.* 103.23 (2009), p. 0238701 (cit. on pp. 9, 24, 27).

[14] P. A. Stokes and P. L. Purdon. “A study of problems encountered in Granger causality analysis from a neuroscience perspective”. In: *Proc. Natl. Acad. Sci. USA* 114.34 (2017), pp. 7063–7072 (cit. on pp. 9, 24, 29).

[15] L. Barnett, A. B. Barrett, and A. K. Seth. “Misunderstandings regarding the application of Granger causality in neuroscience”. In: *Proc. Natl. Acad. Sci. USA* 115.29 (2018), E6676–E6677 (cit. on pp. 9, 29).

[16] L. Barnett, A. B. Barrett, and A. K. Seth. “Solved problems for Granger causality in neuroscience: A response to Stokes and Purdon”. In: *NeuroImage* 178 (2018), pp. 744–748 (cit. on pp. 9, 29).

[17] P. A. Stokes and P. L. Purdon. “Correction for Stokes and Purdon, A study of problems encountered in Granger causality analysis from a neuroscience perspective”. In: *Proc. Natl. Acad. Sci. USA* 115.29 (2018), E6964–E6964 (cit. on pp. 9, 29).

[18] L. Barnett and A. K. Seth. “Granger causality for state-space models”. In: *Phys. Rev. E (Rapid Communications)* 91.4 (2015), 040101(R) (cit. on pp. 9, 24, 30, 32, 45–47, 49, 50).

[19] Paul L Williams and Randall D Beer. “Nonnegative decomposition of multivariate information”. In: *arXiv preprint arXiv:1004.2515* (2010) (cit. on pp. 10, 64, 65, 67, 71, 73, 86, 107, 120, 131, 145, 158, 163, 230).

[20] Lennart Van Hirtum, Patrick De Causmaecker, Jens Goemaere, et al. “A computation of D (9) using FPGA Supercomputing”. In: *arXiv preprint arXiv:2304.03039* (2023) (cit. on p. 12).

[21] Christian Jäkel. “A computation of the ninth Dedekind Number”. In: *arXiv preprint arXiv:2304.00895* (2023) (cit. on p. 12).

[22] Patricia Wollstadt, Sebastian Schmitt, and Michael Wibral. “A Rigorous Information-Theoretic Definition of Redundancy and Relevancy in Feature Selection Based on (Partial) Information Decomposition.” In: *J. Mach. Learn. Res.* 24 (2023), pp. 131–1 (cit. on p. 12).

[23] Paul L Williams and Randall D Beer. “Generalized measures of information transfer”. In: *arXiv preprint arXiv:1102.1507* (2011) (cit. on p. 13).

[24] Holk Cruse and Malte Schilling. “Mental states as emergent properties: From walking to consciousness”. In: *Open Mind*. Open MIND. Frankfurt am Main: MIND Group, 2014 (cit. on p. 14).

[25] Lionel Barnett and Anil K Seth. “Dynamical independence: discovering emergent macroscopic processes in complex dynamical systems”. In: *Physical Review E* 108.1 (2023), p. 014304 (cit. on p. 14).

[26] Fernando E Rosas, Pedro AM Mediano, Henrik J Jensen, et al. “Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data”. In: *arXiv preprint arXiv:2004.08220* (2020) (cit. on pp. 14, 107).

[27] Chris G Langton. “Computation at the edge of chaos: Phase transitions and emergent computation”. In: *Physica D: nonlinear phenomena* 42.1-3 (1990), pp. 12–37 (cit. on p. 14).

[28] Joseph T Lizier. *The local information dynamics of distributed computation in complex systems*. Springer Science & Business Media, 2012 (cit. on p. 14).

[29] Joseph T Lizier, Benjamin Flecker, and Paul L Williams. “Towards a synergy-based approach to measuring information modification”. In: *2013 IEEE Symposium on Artificial Life (ALIFE)*. IEEE. 2013, pp. 43–51 (cit. on p. 14).

[30] David Alexander Ehrlich, Andreas Christian Schneider, Viola Priesemann, Michael Wibral, and Abdullah Makkeh. “A Measure of the Complexity of Neural Representations based on Partial Information Decomposition”. In: *Transactions on Machine Learning Research* (2023) (cit. on pp. 15, 220).

[31] Michael Wibral, Viola Priesemann, Jim W Kay, Joseph T Lizier, and William A Phillips. “Partial information decomposition as a unified approach to the specification of neural goal functions”. In: *Brain and cognition* 112 (2017), pp. 25–38 (cit. on pp. 15–17, 64, 65, 76, 83, 107).

[32] Andy Clark. “Whatever next? Predictive brains, situated agents, and the future of cognitive science”. In: *Behavioral and brain sciences* 36.3 (2013), pp. 181–204 (cit. on p. 16).

[33] Ravid Shwartz-Ziv and Naftali Tishby. “Opening the black box of deep neural networks via information”. In: *arXiv preprint arXiv:1703.00810* (2017) (cit. on p. 16).

[34] Marcel Graetz, Abdullah Makkeh, Andreas C Schneider, et al. “Infomorphic networks: Locally learning neural networks derived from partial information decomposition”. In: *arXiv preprint arXiv:2306.02149* (2023) (cit. on pp. 16, 17, 221).

[35] Robert M Fano. “Transmission of Information: A Statistical Theory of Communication MIT Press”. In: *Cambridge, Mass. and Wiley, New York* (1961) (cit. on pp. 17, 69, 125).

[36] Joseph Lizier and Mikail Rubinov. “Multivariate construction of effective computational networks from observational data”. In: *Preprint of the Max Planck Society* 25 (2012) (cit. on pp. 18, 203).

[37] Lionel Barnett, Suresh D Muthukumaraswamy, Robin L Carhart-Harris, and Anil K Seth. “Decreased directed functional connectivity in the psychedelic state”. In: *NeuroImage* 209 (2020), p. 116462 (cit. on p. 19).

[38] Felipe Llinares-López, Mahito Sugiyama, Laetitia Papaxanthos, and Karsten Borgwardt. “Fast and memory-efficient significant pattern mining via permutation testing”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, pp. 725–734 (cit. on pp. 20, 186, 187, 189, 191, 193, 196, 201, 212, 225).

[39] Felipe Llinares López. “Significant Pattern Mining for Biomarker Discovery”. PhD thesis. ETH Zurich, 2018 (cit. on pp. 20, 225).

[40] Alla Brodski-Guerniero, Marcus J Naumer, Vera Moliadze, et al. “Predictable information in neural signals during resting state is reduced in autism spectrum disorder”. In: *Human brain mapping* (2018) (cit. on pp. 20, 203).

[41] L. Barnett and T. Bossomaier. “Transfer entropy as a log-likelihood ratio”. In: *Phys. Rev. Lett.* 109.13 (2013), p. 0138105 (cit. on p. 24).

[42] J. Neyman and E. S. Pearson. “On the problem of the most efficient tests of statistical hypotheses”. In: *Phil. Trans. R. Soc. A* 231 (1933), pp. 289–337 (cit. on p. 24).

[43] S. S. Wilks. “The large-sample distribution of the likelihood ratio for testing composite hypotheses”. In: *Ann. Math. Stat.* 6.1 (1938), pp. 60–62 (cit. on pp. 24, 29).

[44] A. Wald. “Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large”. In: *T. Am. Math. Soc.* 54.3 (1943), pp. 426–482 (cit. on pp. 24, 41, 44).

[45] M. Ding, Y. Chen, and S. L. Bressler. “Granger Causality: Basic Theory and Application to Neuroscience”. In: *Handbook of Time Series Analysis*. Ed. by B. Schelter, M. Winterhalder, and J. Timmer. Wiley-VCH Verlag GmbH & Co. KGaA, 2006, pp. 437–460 (cit. on pp. 24, 29).

[46] Y. Chen, S. L. Bressler, and M. Ding. “Frequency decomposition of conditional Granger causality and application to multivariate neural field potential data”. In: *J. Neuro. Methods* 150 (2006), pp. 228–237 (cit. on pp. 24, 29).

[47] M. Dhamala, G. Rangarajan, and M. Ding. “Analyzing information flow in brain networks with nonparametric Granger causality”. In: *Neuroimage* 41 (2008), pp. 354–362 (cit. on pp. 24, 30).

[48] M. Dhamala, G. Rangarajan, and M. Ding. “Estimating Granger causality from Fourier and wavelet transforms of time series data”. In: *Phys. Rev. Lett.* 100 (2008), p. 018701 (cit. on pp. 24, 30).

[49] L. Barnett and A. K. Seth. “The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference”. In: *J. Neurosci. Methods* 223 (2014), pp. 50–68 (cit. on pp. 24, 30).

[50] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. 3rd. New York, NY, USA: Springer Science+Business Media, LLC, 2005 (cit. on pp. 24, 30, 31, 49, 55).

[51] V. Solo. “State-space analysis of Granger-Geweke causality measures with application to fMRI”. In: *Neural Comput.* 28.5 (May 2016), pp. 914–949 (cit. on pp. 24, 29, 30, 46, 50).

[52] J. Doob. *Stochastic Processes*. New York: John Wiley, 1953 (cit. on p. 25).

[53] P. Masani. “Recent Trends in Multivariate Prediction Theory”. In: *Multivariate Analysis*. Ed. by P. R. Krishnaiah. New York: Academic Press, 1966, pp. 351–382 (cit. on p. 25).

[54] Yu. A. Rozanov. *Stationary Random Processes*. San Francisco: Holden-Day, 1967 (cit. on p. 25).

[55] P. Whittle. “On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix”. In: *Biometrika* 50.1,2 (1963), pp. 129–134 (cit. on p. 26).

[56] E. J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems*. Philadelphia, PA, USA: SIAM, 2012 (cit. on pp. 26, 49).

[57] N. Wiener and P. Masani. “The prediction theory of multivariate stochastic processes: I. The regularity condition”. In: *Acta Math.* 98 (1957), pp. 111–150 (cit. on p. 26).

[58] G. T. Wilson. “The factorization of matricial spectral densities”. In: *SIAM J. Appl. Math.* 23.4 (1972), pp. 420–426 (cit. on p. 26).

[59] L. Barnett and A. K. Seth. “Behaviour of Granger causality under filtering: Theoretical invariance and practical application”. In: *J. Neurosci. Methods* 201.2 (2011), pp. 404–419 (cit. on p. 28).

[60] G. Buzsáki and A. Draguhn. “Neuronal oscillations in cortical networks”. In: *science* 304.5679 (2004), pp. 1926–1929 (cit. on p. 28).

[61] A. D. R. McQuarrie and C.-L. Tsai. *Regression and Time Series Model Selection*. Singapore: World Scientific Publishing, 1998 (cit. on pp. 29, 43).

[62] L. Faes, S. Stramaglia, and D. Marinazzo. "On the interpretability and computational reliability of frequency-domain Granger causality". In: *F1000Research* 6.1710 (2017). Version 1; Referees: 2 approved (cit. on p. 29).

[63] M. Dhamala, H. Liang, S. L. Bressler, and M. Ding. "Granger-Geweke causality: Estimation and interpretation". In: *NeuroImage* 175 (2018), pp. 460–463 (cit. on p. 29).

[64] J.-M. Dufour and T. Taamouti. "Short and long run causality measures: Theory and inference". In: *J. Econometrics* 154.1 (2010), pp. 42–58 (cit. on p. 30).

[65] A. A. Mohsenipour. "On the Distribution of Quadratic Expressions in Various Types of Random Vectors". PhD thesis. Electronic Thesis and Dissertation Repository, 955.: The University of Western Ontario, Dec. 2012 (cit. on pp. 31, 47).

[66] J. D. Hamilton. *Time Series Analysis*. Princeton, NJ: Princeton University Press, 1994 (cit. on p. 32).

[67] H. Lütkepohl. "Testing for causation between two variables in higher dimensional VAR models". In: *Studies in Applied Econometrics*. Ed. by H. Schneeweiß and K. Zimmerman. Heidelberg: Physica-Verlag HD, 1993, pp. 75–91 (cit. on p. 32).

[68] E. J. Hannan and B. G. Quinn. "The Determination of the Order of an Autoregression". In: *J. Roy. Stat. Soc. B Met.* 41.2 (1979), pp. 190–195 (cit. on p. 43).

[69] J.-M. Dufour and D. Pelletier. "Practical methods for modeling weak VARMA processes: Identification, estimation and specification with a Macroeconomic application". In: *J. Bus. Econ. Stat.* 40.3 (2022), pp. 1140–1152 (cit. on p. 44).

[70] P. P. Mitra and H. Bokil. *Observed Brain Dynamics*. New York: Oxford University Press, 2008 (cit. on p. 46).

[71] A. Klein, G. Mélard, and T. Zahaf. "Construction of the exact Fisher information matrix of Gaussian time series models by means of matrix differential rules". In: *Linear Algebra Appl.* 321.1 (2000). Eighth Special Issue on Linear Algebra and Statistics, pp. 209–232 (cit. on p. 47).

[72] D. A. Jones. "Statistical analysis of empirical models fitted by optimization". In: *Biometrika* 70.1 (Apr. 1983), pp. 67–88 (cit. on p. 47).

[73] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge University Press, 2000 (cit. on p. 49).

[74] A. B. Barrett, L. Barnett, and A. K. Seth. "Multivariate Granger causality and generalized variance". In: *Phys. Rev. E* 81.4 (2010), p. 041907 (cit. on p. 59).

[75] M. Studený and J. Vejnarová. "The Multiinformation Function as a Tool for Measuring Stochastic Dependence". In: *Learning in Graphical Models*. Ed. by Michael I. Jordan. Dordrecht: Springer Netherlands, 1998, pp. 261–297 (cit. on p. 59).

[76] Naama Brenner, William Bialek, and Rob de Ruyter Van Steveninck. "Adaptive rescaling maximizes information transmission". In: *Neuron* 26.3 (2000), pp. 695–702 (cit. on p. 64).

[77] Peter E Latham and Sheila Nirenberg. "Synergy, redundancy, and independence in population codes, revisited". In: *Journal of Neuroscience* 25.21 (2005), pp. 5195–5206 (cit. on p. 64).

[78] Adam A Margolin, Ilya Nemenman, Katia Basso, et al. "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context". In: *BMC bioinformatics*. Vol. 7. Springer. 2006, S7 (cit. on p. 64).

[79] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, and Jürgen Jost. "Shared information—New insights and problems in decomposing information in complex systems". In: *Proceedings of the European conference on complex systems 2012*. Springer. 2013, pp. 251–269 (cit. on pp. 64, 227).

[80] Malte Harder, Christoph Salge, and Daniel Polani. "Bivariate measure of redundant information". In: *Physical Review E* 87.1 (2013), p. 012130 (cit. on pp. 64, 86).

[81] Rick Quax, Omri Har-Shemesh, and Peter Sloot. "Quantifying synergistic information using intermediate stochastic variables". In: *Entropy* 19.2 (2017), p. 85 (cit. on pp. 64, 66).

[82] Paolo Perrone and Nihat Ay. "Hierarchical quantification of synergy in channels". In: *Frontiers in Robotics and AI* 2 (2016), p. 35 (cit. on pp. 64, 174).

[83] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. "Quantifying unique information". In: *Entropy* 16.4 (2014), pp. 2161–2183 (cit. on pp. 64, 81, 82, 86, 142, 158, 159, 175).

[84] Joseph T Lizier, Mikhail Prokopenko, and Albert Y Zomaya. “Local measures of information storage in complex distributed computation”. In: *Information Sciences* 208 (2012), pp. 39–54 (cit. on pp. 64, 84).

[85] Thomas Schreiber. “Measuring information transfer”. In: *Physical review letters* 85.2 (2000), p. 461 (cit. on pp. 64, 84).

[86] Michael Wibral, Nicolae Pampu, Viola Priesemann, et al. “Measuring information-transfer delays”. In: *PloS one* 8.2 (2013), e55809 (cit. on pp. 64, 84).

[87] Joseph T Lizier, Mikhail Prokopenko, and Albert Y Zomaya. “Local information transfer as a spatiotemporal filter for complex systems”. In: *Physical Review E* 77.2 (2008), p. 026110 (cit. on pp. 64, 84).

[88] Jim W Kay and WA Phillips. “Coherent Infomax as a computational goal for neural systems”. In: *Bulletin of mathematical biology* 73.2 (2011), pp. 344–372 (cit. on pp. 64, 65, 83, 106).

[89] Michael Wibral, Joseph T Lizier, and Viola Priesemann. “Bits from brains for biologically inspired computing”. In: *Frontiers in Robotics and AI* 2 (2015), p. 5 (cit. on pp. 65, 68, 106).

[90] Gustavo Deco and Bernd Schürmann. *Information dynamics: foundations and applications*. Springer Science & Business Media, 2012 (cit. on p. 65).

[91] Conor Finn and Joseph Lizier. “Pointwise partial information decomposition using the specificity and ambiguity lattices”. In: *Entropy* 20.4 (2018), p. 297 (cit. on pp. 65, 74, 75, 81, 84–87, 93, 94, 107, 124–126, 134, 158).

[92] Aaron J Gutknecht, Michael Wibral, and Abdullah Makkeh. “Bits and Pieces: Understanding Information Decomposition from Part-whole Relationships and Formal Logic”. In: *arXiv preprint arXiv:2008.09535* (2020) (cit. on pp. 67, 71, 94).

[93] Conor Finn and Joseph Lizier. “Probability Mass Exclusions and the Directed Components of Mutual Information”. In: *Entropy* 20.11 (2018), p. 826 (cit. on pp. 68–70, 74, 75, 87, 144).

[94] Abdullah Makkeh, Dirk Oliver Theis, and Raul Vicente. “Bivariate partial information decomposition: The optimization perspective”. In: *Entropy* 19.10 (2017), p. 530 (cit. on p. 76).

[95] Abdullah Makkeh and Dirk Oliver Theis. “Optimizing Bivariate Partial Information Decomposition”. In: *arXiv preprint arXiv:1802.03947* (2018) (cit. on p. 76).

[96] Robin Ince. “Measuring multivariate redundant information with pointwise common change in surprisal”. In: *Entropy* 19.7 (2017), p. 318 (cit. on pp. 76, 81, 82, 86, 158).

[97] Johannes Rauh, Pradeep Banerjee, Eckehard Olbrich, Jürgen Jost, and Nils Bertschinger. “On extractable shared information”. In: *Entropy* 19.7 (2017), p. 328 (cit. on p. 80).

[98] Philip M Woodward and Ian L Davies. “Information theory and inverse probability in telecommunication”. In: *Proceedings of the IEE-Part III: Radio and Communication Engineering* 99.58 (1952), pp. 37–44 (cit. on p. 80).

[99] Andre M Bastos, W Martin Usrey, Rick A Adams, et al. “Canonical microcircuits for predictive coding”. In: *Neuron* 76.4 (2012), pp. 695–711 (cit. on p. 84).

[100] Matthew Larkum. “A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex”. In: *Trends in neurosciences* 36.3 (2013), pp. 141–151 (cit. on p. 84).

[101] Joseph T Lizier, Benjamin Flecker, and Paul L Williams. “Towards a synergy-based approach to measuring information modification”. In: *2013 IEEE Symposium on Artificial Life (ALIFE)*. IEEE. 2013, pp. 43–51 (cit. on pp. 84, 107).

[102] Michael Wibral, Conor Finn, Patricia Wollstadt, Joseph T Lizier, and Viola Priesemann. “Quantifying information modification in developing neural networks via partial information decomposition”. In: *Entropy* 19.9 (2017), p. 494 (cit. on pp. 84, 107).

[103] Patricia Wollstadt, Joseph T Lizier, Raul Vicente, et al. “IDTxl: The Information Dynamics Toolkit xl: a Python package for the efficient analysis of multivariate information dynamics in networks”. In: *arXiv preprint arXiv:1807.10459* (2018) (cit. on pp. 85, 129, 187, 212).

[104] Virgil Griffith and Christof Koch. “Quantifying synergistic mutual information”. In: *Guided Self-Organization: Inception*. Springer, 2014, pp. 159–190 (cit. on pp. 86, 88, 89).

[105] J Crampton and G Loizou. *Embedding a poset in a lattice,*" tech. rep. Tech. Rep. BBKCS-0001, Birkbeck College, University of London, 2000 (cit. on p. 93).

[106] Andrzej P Ruszczyński and Andrzej Ruszczynski. *Nonlinear optimization*. Vol. 13. Princeton university press, 2006 (cit. on p. 95).

[107] Elad Schneidman, William Bialek, and Michael J Berry. “Synergy, redundancy, and independence in population codes”. In: *Journal of Neuroscience* 23.37 (2003), pp. 11539–11553 (cit. on p. 106).

[108] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003 (cit. on p. 106).

[109] Rajesh PN Rao and Dana H Ballard. “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects”. In: *Nature neuroscience* 2.1 (1999), pp. 79–87 (cit. on p. 106).

[110] Johannes Rauh. “Secret sharing and shared information”. In: *Entropy* 19.11 (2017), p. 601 (cit. on p. 107).

[111] Fernando Rosas, Pedro AM Mediano, Martı n Ugarte, and Henrik J Jensen. “An information-theoretic approach to self-organisation: Emergence of complex interdependencies in coupled dynamical systems”. In: *Entropy* 20.10 (2018), p. 793 (cit. on p. 107).

[112] Richard P Stanley. “Enumerative Combinatorics, vol. 1. 1997”. In: *Cambridge Stud. Adv. Math* (1997) (cit. on p. 114).

[113] Peter Tittmann. *Einführung in die Kombinatorik*. Springer, 2014 (cit. on p. 122).

[114] Jason Crampton and George Loizou. “Two partial orders on the set of antichains”. In: *Research note, September* (2000) (cit. on pp. 124, 139, 148, 167).

[115] Abdullah Makkeh, Aaron J Gutknecht, and Michael Wibral. “Introducing a differentiable measure of pointwise shared information”. In: *Physical Review E* 103.3 (2021), p. 032149 (cit. on pp. 124, 125, 129, 144, 158).

[116] Raymond M Smullyan. *First-order logic*. Courier Corporation, 1995 (cit. on pp. 130, 146).

[117] Ryan G James, Jeffrey Emenheiser, and James P Crutchfield. “Unique information via dependency constraints”. In: *Journal of Physics A: Mathematical and Theoretical* 52.1 (2018), p. 014002 (cit. on p. 139).

[118] Nihat Ay, Daniel Polani, and Nathaniel Virgo. “Information decomposition based on cooperative game theory”. In: *arXiv preprint arXiv:1910.05979* (2019) (cit. on p. 139).

[119] Fernando E Rosas, Pedro AM Mediano, Borzoo Rassouli, and Adam B Barrett. “An operational information decomposition via synergistic disclosure”. In: *Journal of Physics A: Mathematical and Theoretical* 53.48 (2020), p. 485001 (cit. on pp. 139, 158, 159, 167, 174, 180).

[120] Cesare Magri. “On shared and multiple information”. In: *arXiv preprint arXiv:2107.11032* (2021) (cit. on p. 158).

[121] Michael Kleinman, Alessandro Achille, Stefano Soatto, and Jonathan C Kao. “Redundant information neural estimation”. In: *Entropy* 23.7 (2021), p. 922 (cit. on p. 158).

[122] David Sigtermans. “A path-based partial information decomposition”. In: *Entropy* 22.9 (2020), p. 952 (cit. on p. 158).

[123] Malte Harder, Christoph Salge, and Daniel Polani. “Bivariate measure of redundant information”. In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 87.1 (2013), p. 12130. arXiv: 1207.2080 (cit. on pp. 158, 159).

[124] Virgil Griffith, Edwin Chong, Ryan James, Christopher Ellison, and James Crutchfield. “Intersection Information Based on Common Randomness”. In: *Entropy* 16.4 (Apr. 2014), pp. 1985–2000 (cit. on p. 158).

[125] Virgil Griffith and Tracey Ho. “Quantifying redundant information in predicting a target random variable”. In: *Entropy* 17.7 (2015), pp. 4644–4653 (cit. on p. 158).

[126] Adam B Barrett. “Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems”. In: *Physical Review E* 91.5 (2015), p. 52802 (cit. on p. 158).

[127] Artemy Kolchinsky. “A Novel Approach to the Partial Information Decomposition”. In: *Entropy* 24.3 (2022), p. 403 (cit. on p. 158).

[128] Ari Pakman, Amin Nejatbakhsh, Dar Gilboa, et al. “Estimating the unique information of continuous variables”. In: *Advances in neural information processing systems* 34 (2021), pp. 20295–20307 (cit. on pp. 158, 175).

[129] Ryan G James, Jeffrey Emenheiser, and James P Crutchfield. “Unique information and secret key agreement”. In: *Entropy* 21.1 (2018), p. 12 (cit. on pp. 158, 175).

[130] Steven J van Enk. “Pooling probability distributions and partial information decomposition”. In: *Physical Review E* 107.5 (2023), p. 054133 (cit. on pp. 158, 167).

[131] Virgil Griffith and Christof Koch. “Quantifying synergistic mutual information”. In: *Guided self-organization: inception*. Springer, 2014, pp. 159–190 (cit. on p. 158).

[132] Aaron J Gutknecht, Michael Wibral, and Abdullah Makkeh. “Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic”. In: *Proceedings of the Royal Society A* 477.2251 (2021), p. 20210110 (cit. on pp. 158, 160, 166, 170, 175).

[133] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, and Jürgen Jost. “Shared information—New insights and problems in decomposing information in complex systems”. In: *Springer Proceedings in Complexity*. Springer, 2013, pp. 251–269. arXiv: 1210.5902 (cit. on p. 159).

[134] David A Ehrlich, Andreas C Schneider, Michael Wibral, Viola Priesemann, and Abdullah Makkeh. “Partial Information Decomposition Reveals the Structure of Neural Representations”. In: *arXiv preprint arXiv:2209.10438* (2022) (cit. on p. 159).

[135] Daniel Chicharro and Stefano Panzeri. “Synergy and redundancy in dual decompositions of mutual information gain and information loss”. In: *Entropy* 19.2 (2017), p. 71 (cit. on pp. 167, 181).

[136] Mahito Sugiyama, Felipe Llinares López, Niklas Kasenburg, and Karsten M Borgwardt. “Significant subgraph mining with multiple testing correction”. In: *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM. 2015, pp. 37–45 (cit. on pp. 186, 187, 189, 191, 193, 201).

[137] Ed Bullmore and Olaf Sporns. “Complex brain networks: graph theoretical analysis of structural and functional systems”. In: *Nature reviews neuroscience* 10.3 (2009), pp. 186–198 (cit. on pp. 187, 188).

[138] Tiago A Schieber, Laura Carpi, Albert Díaz-Guilera, et al. “Quantification of network structural dissimilarities”. In: *Nature communications* 8.1 (2017), pp. 1–10 (cit. on p. 187).

[139] Ahmad Mheich, Mahmoud Hassan, Mohamad Khalil, et al. “SimiNet: a novel method for quantifying brain network similarity”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.9 (2017), pp. 2238–2249 (cit. on p. 187).

[140] Yutaka Shimada, Yoshito Hirata, Tohru Ikeguchi, and Kazuyuki Aihara. “Graph distance for complex networks”. In: *Scientific reports* 6.1 (2016), pp. 1–6 (cit. on p. 187).

[141] RE Tarone. “A modified Bonferroni method for discrete data”. In: *Biometrics* (1990), pp. 515–522 (cit. on p. 193).

[142] Peter H Westfall, S Stanley Young, et al. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Vol. 279. John Wiley & Sons, 1993 (cit. on pp. 193, 213).

[143] Gerhard Hommel and Frank Krummenauer. “Improvements and modifications of Tarone’s multiple test procedure for discrete data”. In: *Biometrics* (1998), pp. 673–681 (cit. on pp. 193, 216).

[144] Jesse Hemerik and Jelle Goeman. “Exact testing with random permutations”. In: *TEST* 27.4 (2018), pp. 811–825 (cit. on p. 195).

[145] Leonardo Novelli, Patricia Wollstadt, Pedro Mediano, Michael Wibral, and Joseph T Lizier. “Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing”. In: *Network Neuroscience* 3.3 (2019), pp. 827–847 (cit. on p. 203).

[146] Leonardo Novelli and Joseph T Lizier. “Inferring network properties from time series using transfer entropy and mutual information: Validation of multivariate versus bivariate approaches”. In: *Network Neuroscience* 5.2 (2021), pp. 373–404 (cit. on p. 203).

[147] Andrea Avena-Koenigsberger, Bratislav Misic, and Olaf Sporns. “Communication dynamics in complex brain networks”. In: *Nature reviews neuroscience* 19.1 (2018), pp. 17–33 (cit. on p. 203).

[148] Aline Viol, Fernanda Palhano-Fontes, Heloisa Onias, et al. “Characterizing complex networks using entropy-degree diagrams: unveiling changes in functional brain connectivity induced by Ayahuasca”. In: *Entropy* 21.2 (2019), p. 128 (cit. on p. 210).

[149] Andrew Zalesky, Alex Fornito, and Edward T. Bullmore. “Network-based statistic: Identifying differences in brain networks”. In: *NeuroImage* 53.4 (2010), pp. 1197–1207 (cit. on p. 210).

[150] Sandrine Dudoit, Mark J van der Laan, and Katherine S Pollard. “Multiple testing. Part I. Single-step procedures for control of general type I error rates”. In: *Statistical Applications in Genetics and Molecular Biology* 3.1 (2004), pp. 1–69 (cit. on p. 213).

[151] Peter H Westfall and James F Troendle. “Multiple testing with minimal assumptions”. In: *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 50.5 (2008), pp. 745–755 (cit. on p. 213).

[152] Arthur J Roth. “Multiple comparison procedures for discrete test statistics”. In: *Journal of statistical planning and inference* 82.1-2 (1999), pp. 101–117 (cit. on p. 215).

[153] John Aitchison. “Principles of compositional data analysis”. In: *Lecture Notes-Monograph Series* (1994), pp. 73–81 (cit. on p. 220).

[154] Kyle Schick-Poland, Abdullah Makkeh, Aaron J Gutknecht, et al. “A partial information decomposition for discrete and continuous variables”. In: *arXiv preprint arXiv:2106.12393* (2021) (cit. on p. 220).

[155] Rudolf Carnap. *Logical foundations of probability*. Vol. 2. Citeseer, 1962 (cit. on p. 222).

[156] Rasmus Rendsvig and John Symons. “Epistemic Logic”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University, 2021 (cit. on p. 227).

[157] Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Vardi. *Reasoning about knowledge*. MIT press, 2004 (cit. on p. 227).

[158] Judea Pearl. *Causality*. Cambridge university press, 2009 (cit. on p. 229).

[159] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017 (cit. on p. 229).

[160] Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. “Quantifying causal influences”. In: (2013) (cit. on p. 231).

[161] William Bialek. *Biophysics: searching for principles*. Princeton University Press, 2012 (cit. on p. 231).

[162] Daniel Polani, Thomas Martinetz, and Jan Kim. “An information-theoretic approach for the quantification of relevance”. In: *Advances in Artificial Life: 6th European Conference, ECAL 2001 Prague, Czech Republic, September 10–14, 2001 Proceedings 6*. Springer. 2001, pp. 704–713 (cit. on p. 232).

[163] Artemy Kolchinsky and David H Wolpert. “Semantic information, autonomous agency and non-equilibrium statistical physics”. In: *Interface focus* 8.6 (2018), p. 20180041 (cit. on p. 233).

List of Figures

List of Tables