# Information Theory for Complex Systems Scientists:
# What, Why, & How?

Thomas F. Varley[1,2,*]

[1]*Dept. of Psychological & Brain Sciences, Indiana University, Bloomington, IN, 47405*

[2]*School of Informatics, Computing & Engineering,*

*Indiana University, Bloomington, IN, 47405*

## Abstract

In the 21st century, many of the crucial scientific and technical issues facing humanity can be understood as problems associated with understanding, modelling, and ultimately controlling *complex systems*: systems comprised of a large number of non-trivially interacting components whose collective behaviour can be difficult to predict. Information theory, a branch of mathematics historically associated with questions about encoding and decoding messages, has emerged as something of a *lingua franca* for those studying complex systems, far exceeding its original narrow domain of communication systems engineering. In the context of complexity science, information theory provides a set of tools which allow researchers to uncover the statistical and effective dependencies between interacting components; relationships between systems and their environment; mereological whole-part relationships; and is sensitive to non-linearities missed by commonly parametric statistical models.

In this review, we aim to provide an accessible introduction to the core of modern information theory, aimed specifically at aspiring (and established) complex systems scientists. This includes standard measures, such as Shannon entropy, relative entropy, and mutual information, before building to more advanced topics, including: information dynamics, measures of statistical complexity, information decomposition, and effective network inference. In addition to detailing the formal definitions, in this review we make an effort to discuss how information theory can be *interpreted* and develop the intuition behind abstract concepts like "entropy," in the hope that this will enable interested readings to understand what information is, and how it is used, at a more fundamental level.

* tvarley@iu.edu

## CONTENTS

## I. WHAT ARE COMPLEX SYSTEMS?

If the first decades of the 21th century have provided any insight into what the scientific challenges of tomorrow will be, it is that many of the core problems are tied together by a common thread: complexity emerging from highly-interconnected systems. From climate change, to pandemics, from disinformation to economic inequality, a defining characteristic of many issues facing the modern world is that they deal with the problem of modeling, predicting, and controlling *emergent* properties of systems comprised of thousands, or millions of interacting elements. In the case of pandemics, those elements are individual humans and animals, milling and seething across the globe. They spread germs, get sick, get better, and occasionally die. In the case of "fake news", the elements form an ecosystem of humans and ever-more human-like artificial intelligences, all navigating online environments controlled by increasingly opaque algorithms that were trained on tremendous amounts of data. Climate change threatens the stability of a huge number of interconnected, complex systems, from coral reef ecosystems (which bleach in the face of sustained high temperatures [1]), to agriculture and food-production systems (possibly leading to food shocks, and in extreme cases, famine [2]).

These problems of complexity are interconnected as well: far from being restricted to the disciplinary silos of academia, complex systems can involve many different fields of science and require seeing connections across domains. For example, it takes tremendous amounts of energy to train the artificial intelligences that increasingly dominate our online lives. In an economy run on fossil fuel, this means that the systems driving one phenomenon ("fake news", social media bubbles, "infodemics" [3]) are simultaneously exacerbating another one (climate change) with their large carbon footprints [4]. These simple examples highlight some of the fundamental challenges of modern complex systems science:

1. The systems that make up our world are fundamentally *interconnected* in ways that defy the typical breakdown of academic science disciplines. It is not often that climatologists collaborate with machine learning engineers, although as awareness of the carbon costs of computation increases this is starting to change [4].

2. All of the systems described above are highly non-linear [5], exhibiting complex phenomena such as critical phase changes [6], cascading failures [7], extreme ranges of

intensities [8], and self-organization [9] (to name but a few of the interesting properties complex systems can display).

3. These systems are *multiscale:* at the fundamental level, they involve millions, or even billions of moving parts: modelling the entire structure would require astronomical amounts of data, computing power, and very detailed mathematical and/or algorithmic models of behaviour. Consequently, the ability to think critically about the distinction between "wholes" and "parts" is key [10]. So is understanding when it is possible to reduce the dimensionality with things like coarse-graining [11]. Furthermore, "higher-order" or "synergistic" effects present only at the macro-scale can complicate analyses when systems don't neatly decompose into sums of parts [12].

4. The processes that make up the world are *dynamic:* the past informs on the future and individual moments cannot be assumed to be decoupled from the history that lead up to that point [13].

What might be called "classical" statistical modelling, which makes assumptions of linearity, reductionism, and independence can struggle when attempting to disentangle complex systems. Warren Weaver, a pioneer of early complex system science described a schema with which systems can be categorized:

*"Simple" systems*, which are comprised of a small number of interacting components that can be easily described using mathematical models. For example, Newtonian physics does an excellent job describing the collision between two objects moving through space.

*"Disorganized" complex systems*, which are comprised of a large number of largely non-interacting components and are amenable to aggregation and linear modelling. For example, there are billions of atoms in a gas, but their interactions are trivial and the logic of statistical mechanics describes their (macro-scale) behaviour well.

*"Organized" complex systems*, which make up almost all of the really interesting systems in nature. These systems are composed of large numbers of parts (as in the disorganized systems), but the interactions are non-trivial, with complex correlations and patterns spanning multiple scales.

Organized complex systems, in some sense, combine the worst (or best) of both worlds. The wicked combination of a large number of elements, coupled by non-trivial interactions results in systems that are not particularly amenable to the simple and elegant analytic models of the past. Instead, the only way forward is to tackle the complexity head on, using computer simulations, structure learning, and large models [14].

What, then, is the way forward? Science in the $21^{st}$ century needs a toolkit that can be applied to as many complex systems as possible, both to understand their inner workings, and facilitate interdisciplinary interactions between fields that may never have had much reason to come into contact with each-other. Furthermore, rather than running from the nonlinear, hard-to-model nature of complex data, the science of tomorrow must be able to tackle these systems without throwing important information under the rug by making over-simplifying assumptions of linearity, independence, or equilibrium conditions. Increasingly, the emerging answer to this is "information theory," which is poised to become the *lingua franca* of our increasingly complexity-aware scientific enterprise.

## II. WHAT IS INFORMATION?

Information theory was first developed by Claude Shannon, an engineer working at Bell Labs on the problem of how to reliably send a signal from a sender to a receiver over some potentially noisy channel (such as a telegraph line). In a now-famous moment of remarkable genius, Shannon produced an almost the entirety of classical information theory in a single monograph: *A Mathematical Theory of Communication*, first published in 1948 [15]. In this highly practical context, "information" was understood as referring to signals being propagated through a channel. If someone speaks into a telephone receiver, but all the person at the other end hears is static white-noise, the "information" has clearly been lost, while if the voice comes through loud and clear, then information has been transmitted. Shannon provided a mathematically rigorous formal language with which to discuss how one might quantify how much information was transmitted (or lost), and what the fundamental limits on transmission are in noisy environments. We will describe the idea of information in formal detail below, but intuitively, it can be understood in terms of how our uncertainly is resolved by the data made available to us. If all you hear over the phone is static, you remain very uncertain about the message that the sender was trying to communicate, while if the signal is clear and crisp, all that uncertainty is reduced and you know exactly what the message was with total certainty.[1]

The significance of this development was not lost on Shannon's fellow cyberneticians, and information theory quickly generated interest from a broad and interdisciplinary group of scientists, mathematicians, and even philosophers. The question of what information theory's "natural domain" is has been a contentious one ever since. Despite Shannon's initial narrow, and eminently practical focus, during the cybernetic revolution in the mid-20th information theory was quickly applied to a wide range of fields, from sociology to biology [16, 17]. Shannon himself, alarmed at what he saw as a the hasty misapplication of his theories to disciplines where they didn't belong, pushed back, claiming that, by the mid-1950s that information theory had become a "bandwagon" and was been "oversold" by overeager researchers [18]. Despite Shannon's apparent misgivings, however, in the modern day, information theory has retained its popularity, finding applications in diverse fields such

---

[1] It is not hyperbolic to say that this work formed one of the fundamental foundations on which our modern, digital world was built (on par with the development of electricity, or the internal combustion engine in terms of impact).

as neuroscience [19], ecology [20, 21], artificial intelligence [22], climatology [23], and deep connections have been found between information theoretic formalisms and fundamental physical models such as thermodynamics [24, 25].

It is interesting to consider the question of how it came to be that Shannon's theory, developed within the practical confines of communications engineering, has been able to find such widespread success so far beyond its original home. Has the bandwagon simply continued to run amok, heedless of Shannon's 1956 warning? We would (perhaps unsurprisingly) argue "no" (it's hard to imagine why we would have written this manifesto otherwise). Rather, we claim that information theory's native domain is *not just* the field of communication engineering where it was born but rather **information theory is better understood as being fundamentally about the general mathematics of *inference* [26], particularly, *inference under conditions of uncertainty.*** The original applications of information theory (how to reliably transmit a message from a sender to a receiver along a noisy channel) is itself a particular kind of inference problem, and more a special case of a general framework. When the receiver receives a message transmitted in the presence of noise, the problem of decoding that message is one of making inferences about the most probable contents of the message ("what was the sender likely saying"). Similarly, when engineering the a usable communication channel, the goal is to give the receiver the best possible chance of correctly inferring the senders message. *It is the logic of inference (encoding, decoding, probability and statistics) that informs the development of communications channels, not the other way around.*

This perspective of "information theory as inference" resolves another long-standing tension in the field. Information theory is agnostic to the question of the "meaning" of a message: it never mattered to Shannon what message was being sent over a noisy channel, only the statistics of the symbols that make it up. This distinction has been referred to as the difference between the *syntactic* information (which information theory is concerned with), and the *semantic* information (which it is not) [27]. As long as information theory has been understood as a "mathematical theory of communication", the question of meaning (semantic content) is a very natural one: everyone cares more about the meaning of a message than the statistics of the symbols used to transmit it [2] and information theory's failure to address this has lead some to doubt its appropriateness as a theory of communication.

---

[2] Except engineers, of course

This paper is written with perspective that focusing on the "communication" angle is a mistake, or at least, insufficiently general. When considering "information theory as inference", it is clear that the question of semantic information is largely irrelevant: we are not interested in what a message might mean, but rather, ensuring that we have inferred the correct message from some space of possible messages. The same mathematics, however can apply to any issue where we are attempting to increase our degree of certainty about the state of some as-yet-unknown variable. Whether it is a message coming down a pipe, the future state of an evolving system, or the degree of mutual coupling between two variables. These are the core questions that go into modelling complex systems.

*A Note to the Mathematically Nervous*

It is impossible to write a comprehensive review of information theory (a field of mathematics in it's own right) in complex systems (a very quantitative meta-field) without relying on a certain amount of mathematics. However, throughout this review we have done our best to bolster the formulae with discussions about the intuition behind the measures and how they might be interpreted. Complex systems draws interest from many fields, and many (especially students just discovering the interest for the first time) may not have a comprehensive mathematics background. To paraphrase Albert Einstein, we have tried to "make everything as simple as possible, but no simpler". For readers who wish the refresh on the basics on probability theory (on which information theory is fundamentally based), see the Supplementary Information).

With that caveat in mind, we now turn to the most fundamental building block of information theory: the entropy.

## A. Entropy

The fundamental questions information theory equips us to ask (and answer) are:

1. How uncertain am I about the state of some variable?

2. How much "work" (for want of a better phrase) do I need to do to become *more certain* than I am now?

It is important to note that these two core questions are not (on the surface) actually about information, but rather, they are about *uncertainty*, and how uncertainty changes. This forms the core of Shannon's great insight: information is the reduction of uncertainty associated with observing data. To even begin to mathematically understand information, we must first understand uncertainty. This is done with the "entropy", a mathematical function that quantifies how uncertain we are about the state of some variable.

There is a (likely apocryphal) story that, upon developing his theory, Claude Shannon approached the famous physicist and fellow cybernetician John von Neumann and asked what he should call his new measure. Von Neumann (apparently) replied: *"you should call it entropy...since no one really knows what entropy is, you'll always have the upper hand in a debate."* (Paraphrased). While tongue-in-cheek, this story highlights the reality that, despite Shannon's originally narrow focus, the structure he developed is remarkably general and lends itself to many interpretations (and more than a few misinterpretations as well).

Many intuitive understandings of information entropy have been proposed, and many authors have their own ways of introducing it. Here we provide two interpretations that make sense to us, but we encourage interested readers to go find as many explanations of entropy as they can.

*1.    Entropy as Expected Surprise*

To answer Question 1 ("how uncertain am I about the state of something in the world"), we need a measure of uncertainty; a way to quantify exactly how certain we are about whatever it is we are interested in.

To understand what we mean by quantifying "uncertainty", consider the case of a fair die, where the probability of any face coming up is 1/6. If you were gambling on the die, you would have no reason to pick any face over another and consequently *no outcome would be more surprising than any other.* You would be no more shocked to roll a 2 than a 5. This notion of "surprise" as being related to "uncertainty" becomes clearer if we instead imagine that you are a cheater and, at some point, manage to switch the fair die for a loaded one of your own. This tricky die is balanced in such a way that 2 comes up 2/3 of the time, and all other faces come up with probability 1/15. Knowing this probability distribution ahead of time, you will be *less surprised* to see a 2 come up than you would be to see a 2 come up

on the fair die. Similarly, you would be *more surprised* to see any other number turn up on the loaded die than you would to see that same number on the fair die.

Everyone has an intuitive notion that some events are more surprising than others, and that how surprised you are to see something is directly related to how likely you were to make a correct prediction. This formed the basis of Shannon's derivation entropy, and is the first, fundamental mathematical measure we will introduce. Shannon [15] realized that a quantitative measure of surprise (noted as $h()$) should satisfy a set of four basic criteria. For a random variable $X$ that draws states from the support set $\mathcal{X}$ according to the probability distribution $P(x)$:

1. The surprise associated with a particular outcome $X = x$ should be inversely proportional to the probability $P(x)$: the more probable an event, the less surprising it is.

2. Surprise should never be negative. Formally:
   $h(x) \geq 0$ for all $x \in \mathcal{X}$.

3. Events that are guaranteed to happen are totally unsurprising. Formally:
   If $P(x) = 1$ than $h(x) = 0$.

4. Surprise should be additive: our surprise at the outcome of two simultaneous, independent coin flips should be equal to our surprise at the outcome of the same coin flipped two different times. Formally:
   $h(x_1, x_2) = h(x_1) + h(x_2) \iff X_1 \perp X_2$.

He went on to show that the function uniquely specified by these desiderata is:

$$h(x) = \log\left(\frac{1}{P(x)}\right) \tag{1}$$

This value is often referred to as the surprise, Shannon information content, or the local entropy of a specific outcome of random variable $X$. The less probable an outcome, the more surprising it is. We can now ask "on average, how surprised are we as we watch this variable over time?" Rare events are very surprising but we aren't that surprised very often. In contrast, frequent events are unsurprising, so we spend a lot of time not being very surprised. This can be quantified by the *expected surprise* of a distribution $P(X)$:

$$H(X) = \mathbb{E}_X[h(X)] = -\sum_{x \in \mathcal{X}} P(x) \log(P(x)) \tag{2}$$

This is now-famous equation of Shannon entropy: a function that takes in a discrete random variable and returns a measure of, on average, how uncertain are we about the state of that variable. $H(X)$ has a number of appealing mathematical properties: it is maximal when all outcomes are equiprobable. This matches the intuition laid out in the fair dice example above: when all outcomes have the same chance of occurring, we are maximally uncertain about the outcome of the next roll. Furthermore, by convention we say that $0 \log(0) = 0$, so impossible events have no impact on our uncertainty. Knowing that a fair die can never roll a 100 doesn't decrease our uncertainty about which of the six faces will come up next.

### 2. Entropy as Required Information

Another, related, way of thinking about entropy is based on Question 2: "How much work do I need to do to become more certain than I am now?". Forget about surprise for a moment, and ask "on average, what is the minimum number of yes/no questions would I need to determine the number rolled on each dice (the fair and the weighted)?"

In the case of the fair die, you have no reason to suspect any particular face to be more likely than another, so a guess of "it is $x$" will only be right one time in six. You can start whittling it down with clever questions like "is it an odd number," however, on average, it will take you $\approx 2.59$ questions to uniquely determine the state of the die. In the case of the loaded die, however, if you start with "is it 2", you'll be right two times in three. On average, it will only take $\approx 1.7$ guesses to determine the outcome of a roll of the loaded die.

Every time you ask a question, you *rule out* possible outcomes, reducing your uncertainty and *gaining information about the state of the die* (for a fascinating discussion of information as probability mass exclusions, see [28]). In the case of the fair die, you need to ask more questions (that is to say, you need more information) to uniquely specify the state of the die, while in the case of the loaded one, since you already know that it's more likely to be two, you need less information. You come into the game with information already "in your pocket", as it were.

Remarkably, for a random variable $X$, the average number of yes/no questions required to uniquely specify the state of $X$ is given by the entropy function $H(X)$, when the base of the logarithm in Eq. 2 is two.

It is important to understand that entropy itself is *not a measure of information in-and-of-itself.* As detailed above, it is fundamentally a measure of uncertainty. The relationship between uncertainty and information is symmetric: the entropy of a variable doesn't tell you how much information is in that variable, but rather, how much information you would need to uniquely specify its state. One common way people think of entropy is as a measure of the "capacity" of a random variable to disclose information. You can't really communicate that much information in a coin, since it can only be heads or tails. In contrast, you can communicate a huge amount of information in the English alphabet with its 26 characters (and correspondingly higher entropy).

### 3. Joint Entropy

The classic Shannon entropy is formulated to describe the uncertainty associated with a single variable. A defining feature of complex systems, however, is the presence of many interacting elements and a core use of information theory when researching complex systems is understanding how elements constrain and disclose information about each-other. The simplest generalization of entropy is the joint entropy, which treats sets of variables as a coarse-grained "macro"-variable and gives a measure of the uncertainty about the "macro"-state of the whole system.

The joint entropy of two variables $H(X_1, X_2)$ is given by:

$$H(X_1, X_2) = \mathbb{E}_{X_1, X_2}[h(X_1, X_2)] = -\sum_{\substack{x_1 \in \mathcal{X}_1 \\ x_2 \in \mathcal{X}_2}} P(x_1, x_2) \log(P(x_1, x_2)) \tag{3}$$

The generalization to more than two variables is obvious. If and only if variables $X_1$ and $X_2$ are all statistically independent, then the joint entropy is equal to the sum of the marginal entropies:

$$H(X_1, X_2) = H(X_1) + H(X_2) \iff X_1 \perp X_2 \tag{4}$$

In the case of complex systems of interacting variables, this independence condition is

14

rarely satisfied, however the joint and marginal entropies are generally related by the inequality:

$$H(X_1, X_2) \leq H(X_1) + H(X_2) \tag{5}$$

Equation 5 tells us that we can never be *more* uncertain about the states of $X_1$ and $X_2$ when considering them together than if we add our uncertainty about each variable individually. We note this by saying that the entropy function is fundamentally "subadditive." As we will see in Sec. II C, if $H(X_1, X_2) \leq H(X_1) + H(X_2)$, then $X_1$ and $X_2$ have a non-zero mutual information and knowing the state of one necessarily reduces our uncertainty about the state of the other.

The joint entropy formula highlights a key fact about information theory: variables can be "coarse-grained" (not to be confused with the coarse-graining explored in Sec. **??**), and every joint distribution can be cast as a single-variable distribution by aggregating multiple "micro"-scale variables into a single "macro"-scale variable. For example, imagine $X_1$ and $X_2$ are two fair, independent coins. The joint states are given by {00, 01, 10, 11}, however we could easily describe both variables as a single "macro"-variable $\mathbf{X}$ with states {A,B,C,D}. In this case:

$$H(X_1, X_2) = H(\mathbf{X}) \tag{6}$$

More complex configurations are also allowable. For example, if we had four micro-state variables $X_1, X_2, X_3, X_4$, they could either be coarse-grained into $\mathbf{X}$, or we could map $X_1$ and $X_2$ to a macro $\mathbf{X}_\alpha$ and $X_3$ and $X_4$ to macro $\mathbf{X}_\beta$. Once again:

$$H(X_1, X_2, X_3, X_4) = H(\mathbf{X}_\alpha, \mathbf{X}_\beta) \tag{7}$$

This ability to move easily between analysis of "wholes" (eg, $\mathbf{X}$, $\mathbf{X}_\alpha$, and $\mathbf{X}_\beta$), and constituent "parts" (eg. $X_1, X_2, X_3$, and $X_4$) makes information theory very useful when assessing the structure of complex systems, which are typically multi-scale, displaying non-trivial dynamics and structure at a variety of levels in a hierarchy. For example, a group of interacting neurons can be considered at the level of the whole set (e.g. a ganglion on brain region), or at the level of each neuron and how subsets of the group interact with each-other. For an example of this, see Sec. II D 1.

*4. Conditional Entropy*

The conditional entropy quantifies how uncertain we are about the state of variable *given that we have knowledge of another variable.* As with many information theoretic measures, it is best understood in a Bayesian context: our uncertainty about a variable will change depending on what information we have about the context around it. *The entropy of a variable describes uncertainty about the state of a variable and is not a value intrinsic to elements of the real world in the way something like mass is.* Our level of certainty changes depending on what information we have.

For two variables, the conditional entropy $H(X_1|X_2)$ gives us a measure of how uncertain we are about the state of $X_1$ after we account for information from $X_2$. It is defined by:

$$H(X_1|X_2) = \mathbb{E}_{X_1, X_2}[h(X_1|X_2)] = -\sum_{\substack{x_1 \in \mathcal{X}_1 \\ x_2 \in \mathcal{X}_2}} P(x_1, x_2) \log(P(x_1|x_2)) \tag{8}$$

Note that the expectation is taken with respect to the joint probability $P(X_1, X_2)$ rather than the conditional probability $P(X_1|X_2)$. This is because of the presence of two interacting variables. If we wanted to calculate the conditional entropy of $X_1$, given that we see $X_2$ in some particular, fixed state $x_2$, then the entropy formula takes on it's usual pattern:

$$H(X_1|x_2) = \mathbb{E}_{X_1|x_2}[h(X_1|x_2)] = -\sum_{x_1 \in \mathcal{X}_1} P(x_1|x_2) \log(P(x_1|x_2)) \tag{9}$$

However, if we want to know the effect of *all* possible $x_2 \in \mathcal{X}_2$, we need to sum the conditional entropy of each $x_2$, and weight the resulting contribution of each by the probability that we see that particular state of $X_2 = x_2$.

$$H(X_1|X_2) = -\sum_{x_2 \in X_2} P(x_2) H(X_1|X_2 = x_2) \tag{10}$$

This results in a "nested expectation", which is equivalent to Eq. 8. The conditional entropy can also be written in terms of the joint and marginal entropies:

$$H(X_1|X_2) = H(X_1, X_2) - H(X_2) \tag{11}$$

The intuition is that the uncertainty about $X_1$ that is left over when uncertainty about $X_2$ is removed from the joint state of both. As with the joint entropy, if and only if all the

variables are independent then:

$$H(X_1|X_2) = H(X_1) \iff X_1 \perp X_2 \tag{12}$$

In other words, if knowing the state of $X_2$ does nothing to reduce our uncertainty about the state of $X_1$, then they must be independent. In general:

$$H(X_1|X_2) \leq H(X_1) \tag{13}$$

As with the joint entropy inequalities, this is understood as indicating that knowing the state of other variables can only ever *decrease* our uncertainty (or in the case of independent variables, leave it unchanged). There is no context in which knowing information about $X_2$ could "obfuscate" our knowledge about $X_1$ and make us more uncertain.

It is crucial to understand that, like all entropy measures, the conditional entropy returns a measure of uncertainty about a variable. There is a tendency among students to read $H(X_1|X_2)$ as "the amount of information $X_2$ provides about $X_1$", which is an incorrect interpretation. This confuses the conditional entropy with the mutual information.

### B. Relative Entropy

Also known as the Kullback-Liebler Divergence ($D_{KL}$), the relative entropy provides a measure of how our certainty about the state a variable (or set of variables) changes depending on our beliefs (in this case, various probability distributions $P$ and $Q$ we could use to model a random variable $X$). Formally:

$$D_{KL}(P||Q) = \mathbb{E}_{P(X)}[\log\left(\frac{P(X)}{Q(X)}\right)] = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \tag{14}$$

The KL-divergence has a very natural Bayesian interpretation: if $Q(X)$ is our prior distribution on $X$, and $P(X)$ is our posterior distribution inferred given some data, than the KL-Divergence is the information *gained* when we revise our prior beliefs $Q$ to the updated posterior beliefs $P$. Conversely, the KL-Divergence can be thought of as the amount of information *lost* when $Q$ is used to approximate $P$.

**Note**: The term $D_{KL}(P||Q)$ is typically read as "the KL-Divergence from $Q$ to $P$," despite the fact that $P$ is written first in the argument. It can alternately be read as "the

relative entropy of $P$ with respect to $Q$."

The KL-divergence between two distributions can be (naively) thought of as giving a "distance" between the two distributions, although it is important to realize that it is not a true metric. It cannot be assumed that $D_{KL}(P||Q) = D_{KL}(Q||P)$, nor can it be assumed that it satisfies the triangle inequality (i.e. for three distribution $P$, $Q$, and $W$, $D_{KL}(P||W) \not\leq D_{KL}(P||Q) + D_{KL}(Q||W)$)

### 1. Local Relative Entropy

As with the Shannon entropy, the relative entropy measure is an expectation value averaged over every value of $x \in \mathcal{X}$. This expectation can be "unrolled" to get the information associated with every specific value $X$ can take on.

$$d_{KL}(P(x)||Q(x)) = \log \left(\frac{P(x)}{Q(x)}\right) = h_Q(x) - h_P(x) \tag{15}$$

We can see that $d_{KL}(P(x)||Q(x))$ tells us how much more surprised we are to see $X = x$ when we are basing our predictions on the prior distribution of $Q(X)$ compared to when we use the posterior distribution $P(X)$. The full KL-divergence, then, can be thought of as the average, element-wise difference in the surprise for every state $x$ in the two distributions $P$ and $Q$, weighted by the posterior distribution. The full relative entropy is always greater than zero, however, for specific local values, it can be less than zero, indicating that, for these values, we would be less surprised to see them if we were basing our predictions on the prior, rather than the posterior.

## C. Mutual Information

Like the Shannon entropy, the mutual information between two variables $X_1$ and $X_2$ is one of the foundational measures of information theory. Mutual information provides a measure of the statistical dependency between two random variables: how much does knowing the state of $X_1$ reduce our uncertainty about the state of $X_2$ (and vice versa). Recall from Sec. II A 2 that "information" is equivalent to "reduction in uncertainty", so mutual information provides the first real measure of information we have encountered so far.
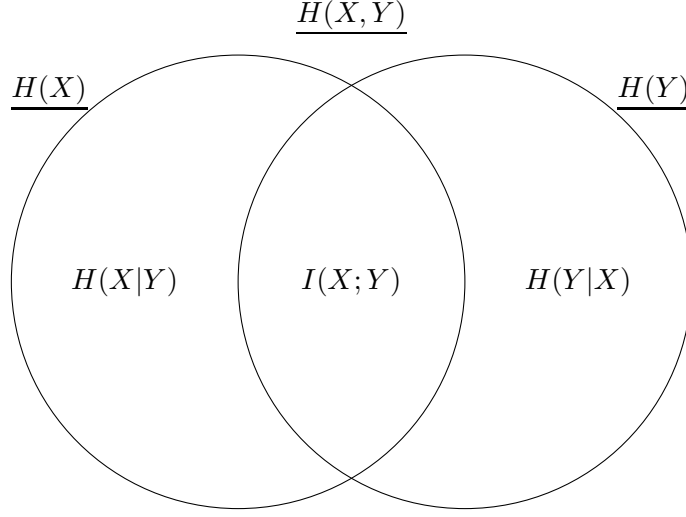
FIG. 1. An entropy diagram showing how the marginal, conditional, and joint entropies are related the mutual information for two interacting variables $X$ and $Y$. The area of each circle corresponds to the *amount of uncertainty* we have about the state of each variable **not** the amount of information we have (this is a common misinterpretation). Each circle corresponds to our total uncertainty about $X$ or $Y$ respectively. The intersection of the Venn diagram is that uncertainty that is common to both variables. If we were to resolve *all* of our uncertainty about $Y$, we would also be resolving some uncertainty about $X$: this overlap is the mutual information $I(X;Y)$. We can also see the conditional entropy is that uncertainty specific to one variable that is left over after resolving the uncertainty specific to the other. Finally, the joint entropy is given by the union of both marginal entropies. From this diagram we can get a visual intuition for the various definitions of mutual information: it's clear that $I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X,Y) - H(Y|X) - H(X|Y)$.

$$I(X_1; X_2) = H(X_1) - H(X_1|X_2) \tag{16}$$

$$= H(X_2) - H(X_2|X_1) \tag{17}$$

$$= H(X_1) + H(X_2) - H(X_1, X_2) \tag{18}$$

$$= H(X_1, X_2) - H(X_1|X_2) - H(X_2|X_1) \tag{19}$$

It can be helpful to break Eq. 16 down into it's component parts. $H(X_1)$ quantifies our initial uncertainty about the state of $X_1$. The conditional entropy $H(X_1|X_2)$ quantifies

how much uncertainty remains *after* learning the state of $X_2$. The difference between these terms is the uncertainty about the state of $X_1$ that is *resolved* by learning the state of $X_2$: the information $X_2$ discloses about $X_1$. Note that while generally $H(X_1) \neq H(X_2)$ and $H(X_1|X_2) \neq H(X_2|X_1)$, the mutual information is symmetric in its arguments (i.e. $I(X_1; X_2) = I(X_2; X_1)$). The other formulations can be derived via simple algebra. See Fig. IIC for a visual aid into the relationships between Eqs. 16, 17, 18, and 19.

The mutual information can also be expressed in terms of the Kullback-Leibler divergence:

$$I(X_1, X_2) = D_{KL}(P(X_1, X_2)||P(X_1) \times P(X_2)) \tag{20}$$

This means that the measure has a natural Bayesian interpretation: in this case, the prior distribution is the product of the marginals: how we would expect the pair of variables to behave if they were independent, with no statistical coupling. The posterior is the true joint probability distribution. The mutual information gives us the *increase in information* we gain when we update our prior (the two variables are uncoupled) to the posterior (the variables are coupled and disclose information about each-other). If $X_1$ and $X_2$ truly are coupled, than knowing the state of one should reduce our uncertainty about the state of the other and vice-versa.

As with the entropy, the mutual information can also be understood as an average measure, formulated as an expectation over the joint state $P(X_1, X_2)$ as seen in Eq. 21:

$$I(X_1; X_2) = \mathbb{E}_{X_1, X_2}\left[\log\left(\frac{P(X_1|X_2)}{P(X_1)}\right)\right] = \sum_{\substack{x_1 \in \mathcal{X}_1 \\ x_2 \in \mathcal{X}_2}} P(x_1, x_2) \log\left(\frac{P(x_1|x_2)}{P(x_1)}\right) \tag{21}$$

The expectation formulation provides another window into the Bayesian interpretation of mutual information. The denominator, $P(X_1)$ can be thought of as our prior beliefs about the distribution of $X_1$. The numerator $P(X_1|X_2)$ is the posterior: our revised beliefs about the distribution of $X_1$ after we have factored in knowledge about the state of $X_2$.

### 1. Local Mutual Information

Like all the other information theoretic measures we have seen so far, mutual information is an expected value, which can be "unrolled", to get the information associated with every possible combination of $X_1$ and $X_2$ values.

$$i(x_1, x_2) = \log\left(\frac{P(x_1|x_2)}{P(x_1)}\right) \tag{22}$$

$$= \log\left(\frac{P(x_1, x_2)}{P(x_1) \times P(x_2)}\right) \tag{23}$$

As with local relative entropy, this measure can be positive or negative. If $P(x_1|x_2) > P(x_1)$, then $i(x_1, x_2) > 0$. Conversely, if $P(x_1|x_2) < P(x_1)$, then $i(x_1, x_2) < 0$. This second case can be thought of as "misinformation:" it appears that the particular joint state we are observing would be *more likely* if $X_1$ and $X_2$ are independent than if they are coupled. Conveniently, the local mutual information relates to the local entropies in all the same ways as the expected mutual information relates to the expected entropies:

$$i(x_1, x_2) = h(x_1) + h(x_2) - h(x_1, x_2) \tag{24}$$

$$= h(x_1, x_2) - h(x_1|x_2) - h(x_2|x_1) \tag{25}$$

$$= h(x_1) - h(x_1|x_2) \tag{26}$$

The local mutual information is not as well-known as the global expected value, however it has been used in a number of different contexts, such as computational chemistry [29], natural language processing [30], network science [31], and image segmentation [32]. In time-series analysis, the local mutual information can be used to construct an "edge-timeseries" [33, 34], which provides a measure of instantaneous coupling between two processes (for example, when are two processes behaving "is if" they were coupled vs. "is if" they were independent).

### 2. Conditional mutual information

As with entropy, the amount of information one element of a system discloses about another can change depending on what information is available from a third (or more) variables. The conditional mutual information between $X_1$ and $X_2$, given $X_3$ is given as:

$$I(X_1; X_2|X_3) = H(X_1|X_3) + H(X_2|X_3) - H(X_1, X_2|X_3) \tag{27}$$

$$= H(X_1|X_3) - H(X_1|X_2, X_3) \tag{28}$$

$$= H(X_1, X_3) + H(X_2, X_3) - H(X_1, X_2, X_3) - H(X_3) \tag{29}$$

Equation 27 shows that the conditional mutual information can be thought of as a measure of the degree to which knowing the state of $X_1$ resolves uncertainty about $X_2$ when information from $X_3$ has been accounted for. This implies that information is distributed over three (or more) variables in a way that constrains how the relationship between $X_1$ and $X_2$ is interpreted.

There are two contexts in which this can occur: when information is *redundantly* shared over multiple variables and when information is *synergistically* present in the joint states of multiple variables. In both cases, examining only the pairwise relationships between elements provides an inaccurate picture of the system under study. Generally, redundant information will result in over-estimates of the extent to which elements are interacting, while synergistic information will result in under-estimates of th same. As we will discuss in Sections V A and V B, this can significantly complicate the problem of constructing network models of complex systems from data.

**Redundant (Shared) Information**

To understand redundant information, consider the case where $X_1$ and $X_2$ both "listen to" $X_3$, but have no coupling between each-other (no information passes from $X_1$ to $X_2$, but they both receive information from $X_3$). Here we have a "common driver effect": even though there is no coupling between $X_1$ and $X_2$, we find that $I(X_1; X_2) > 0$ bit because both are evolving based on a shared upstream input. If we were trying to infer something like a functional connectivity network, we would erroneously put a link between the pair. However, if we condition the mutual information between $X_1$ and $X_2$ on $X_3$, we find that: $I(X_1; X_2|X_3) = 0$ bit.

Why does this happen?

Consider Eq. 27: all the entropy terms have become conditional: if $X_1$ and $X_2$ are both "listening" to the upstream element $X_3$, then knowing the state of $X_3$ resolves uncertainty about the states of both $X_1$ and $X_2$: $H(X_1|X_3) \rightarrow 0$ bit and $H(X_2|X_3) \rightarrow 0$ bit. Furthermore, because $X_3$ is a common driver of both $X_1$ and $X_2$, conditioning the joint states of both downstream elements on $X_3$ results in $H(X_1, X_2|X_3) \rightarrow 0$ because $X_3$ is providing the *same* information to both child elements. Whatever intrinsic uncertainty is left in $X_1$ after all the information from $X_3$ is accounted for won't be resolved by knowledge about

22

$X_2$: at this point, the redundant information has been conditioned out, and since there's no connection between the two, there won't be any statistical dependency.

Here, the process of conditioning the mutual information between two variables and a third is very similar to the process of "regressing out" or "controlling for" a variable in more standard statistical analyses, although as usual, there are no assumptions about linearity or Gaussianity being made here.

**Synergistic Information**

Synergistic information is a little bit more complicated than redundant information, although toy models can help build intuition. Defined generally, synergy referrers to information that is only present in the joint-states of two or more elements of the system and cannot be extracted from single variables considered alone.

Consider three variables, $X_1$, $X_2$, and $X_3$, where $X_1$ and $X_2$ are both random, maximally entropic binary variables, and $X_3 = XOR(X_1, X_2)$ (for the lookup table for the logical XOR gate, see Appendix). Despite the fact that $X_3$ is a deterministic function of both $X_1$ and $X_2$, both $I(X_1; X_3) = 0$ bit and $I(X_2; X_3) = 0$ bit. Due to the synergistic nature of the XOR function, the relationship between all three variables only comes into view when information from all three are accounted for: $I(X_1; X_3|X_2) = 1$ bit. This occurs because the output of the XOR function cannot be known from a single argument: knowledge about both $X_1$ and $X_2$ are necessary to determine whether $XOR(X_1, X_2) = 0$ or 1.

The somewhat counter-intuitive result is that attempts at pairwise inference of connections between $X_1$, $X_2$, and $X_3$ will all return values of 0, however by conditioning we can find that a sort of hyper-edge exists between $X_1, X_2$ and $X_3$: it is only when both $X_1$ and $X_2$ are considered together that we can gain information about $X_3$. Individually, they are completely uninformative.

As with the standard, bivariate mutual information, the local conditional mutual information can also be written out in terms of local joint and conditional entropies in a manner analogous to Eqs. 27, 28, and 29.
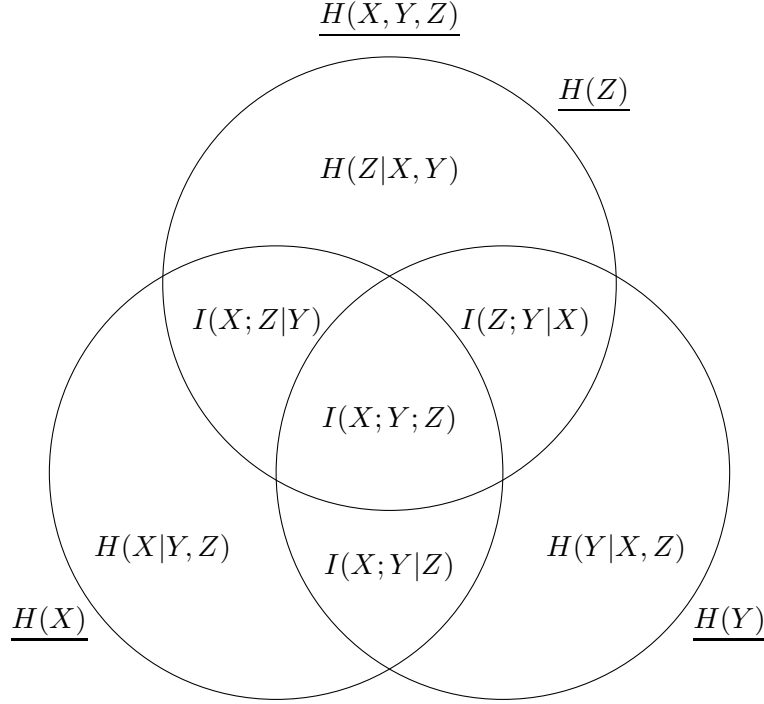
FIG. 2. This Venn diagram generalizes the relationships introduced in Figure II C. However, care should be taken when considering entropy diagrams for more than two variables: the innermost intersection ($I(X;Y;Z)$, corresponding to the co-information or interaction information) is *not* strictly positive [35]. The sign refers to whether the trivariate relationship is redundancy or synergy dominated [36], discussed below.

## D.  Multivariate Generalization of Mutual Information

A defining feature of complex systems, however, is that they are comprised of a large number of interacting elements, often numbering in the hundreds or thousands. This has lead to a natural interest in information theoretic tools that can assess the structure of many-element systems and how they can interact. Arguably the most straight-forward attempts have been multivariate generalizations of mutual information, although this turns out to not be as straightforward as one might think.

Several attempts have been made to generalize the notion of mutual information to more than two variables, although different intuitions can lead to different formalisms. For example, should multivariate mutual information refer to only that information which is redundantly present in *all* $N$ elements, or should it refer to *any* information that is present

in at least two or more elements jointly? While these are identical concepts in the bivariate case, as more variables are added, the situation becomes more complex and different generalizations will quantify different "notions" of shared information.

### 1. Total Correlation

The total correlation [37], also known as multiinformation [38], integration [39] is a multivariate mutual information that generalizes the notion of mutual information as the KL-Divergence between the joint and the independent distributions of all the elements.

$$TC(\mathbf{X}) = D_{KL}(P(X_1, X_2, ...X_N)||P(X_1) \times P(X_2)...P(X_N)) \tag{30}$$

The total correlation is also the natural multivariate generalization of Eq. 18:

$$TC(\mathbf{X}) = \left( \sum_{i=0}^{N} H(X_i) \right) - H(\mathbf{X}) \tag{31}$$

Examining Eq. 31 can provide insights into the behaviour of the total correlation function. $TC(\mathbf{X})$ will be high when all of the individual elements are maximally entropic (that is to say, all possible states appear equiprobably) but the joint state of the system has a low entropy. Said differently, total correlation is high when every individual element explores it's entire state space, but the joint state of the system is restricted to a small subset of the joint state-space.

A good example of this would be synchronization (itself a core topic in complex systems research [40, 41]): imagine a system comprised to 10 elements ($\mathbf{X} = \{X_1, X_2, ...X_{10}\}$), each of which can be in 8 states ($\{S_1, S_2, ...S_8\}$). If all the elements are synchronized and cycling through the sequence of states at the same time, then the entropy of each element will be 3 bit, and the sum of the entropies of all channels will be 30 bit. When considering the joint entropy of the system we find that, in contrast to the individual elements, each of which explores it's full state-space, the system as a whole is restricted to just 8 joint-states out of a possible $8^{10}$ configurations (since every observed joint state is just $S_i$ repeated 10 times). Consequently, the joint entropy is also 3 bit, giving a final total correlation of 27 bit. In contrast, the total correlation would be low when every element is independent of every other, in which case the sum of the individual entropies would be equal to the joint entropy.

Since the total correlation is an extension of the the KL-divergence between the joint distribution and the product of the marginals, it lends itself to the same intuitive Bayesian interpretation as bivariate mutual information. If our prior is that all elements of the system are disconnected, then we would expect $H(\mathbf{X}) >> H(X_i)$, however, in the case of our totally synchronized elements, this would be incorrect and the *increase* in our ability to predict the joint state of the system is indicated by the high total correlation value. That increase in predictive power is a result of the fact that the system explores a very small part of the joint state-space (8 configurations as opposed to $8^{10}$), and so we have far less uncertainty about the joint state at any given time.

### 2. *Dual Total Correlation*

The dual total correlation [42], also known as the binding entropy [43] is another generalization of mutual information to more than two interacting elements. It is based on a generalization of Eq. 19:

$$DTC(\mathbf{X}) = H(\mathbf{X}) - \sum_{i=0}^{N} H(X_i|\mathbf{X}_{-i}) \tag{32}$$

Where $\mathbf{X}_{-i}$ indicates the set of all elements *excluding* the *ith* element. The conditional entropy term $H(X_i|\mathbf{X}_{-i})$ quantifies the uncertainty in the $i^{th}$ element that cannot be resolved with information from any other combination of elements in the system: the uncertainty that is intrinsic to that element in particular. This is often referred to as the residual entropy [43]. The dual total correlation gives the amount of information left over when the intrinsic information associated with each individual element is removed from the joint entropy of the whole: that is to say, all the information that is shared by two or more elements of the system.

Unlike the total correlation, which monotonically increases with the amount of dependency between the elements, the dual total correlation has the interesting property of being low both when all the elements of the system are totally coupled (as in the synchronization example given above), and totally independent. In the case of the 10-element, totally synchronized system described above, the dual total correlation is equivalent to the joint entropy (3 bit), since there is no residual entropy specific to any element (the state of any part of the system can be predicted with total accuracy with knowledge of any other element, since

26

they're always all the same). Similarly, if all the elements are independent, then the joint entropy is high ($\approx \log(8^{10})$), but so is the sum of all the residual entropies (since, by definition, information about the $i^{th}$ element cannot be extracted from independent variables).

The dual total correlation is high in an intermediate zone where the joint entropy of the whole system is high, but the residual entropy of every element is low (i.e. the state of any individual element can be predicted with a high degree of certainty with information present in some combination of other elements). The occurs when complex, informative dependencies exist between the elements in a way that doesn't overly-constrain the flexibility of the whole.

### 3.  Co-Information

Also known as the interaction information [44], the co-information is a multivariate generalization of mutual information that attempts to preserve the notion that mutual information is that information present in *all* the interacting variables. Contrast this with the total correlation (which focuses on mutual information as the $D_{KL}$ between the joint and independent distributions) and the dual total correlation (which aggregates all the information not specific to a single variable) [26, 45].

Based on the diagram in Fig. II D, we can see that, for three variables, the $C(X;Y;Z)$ is the information that is present in all three variables and can be calculated as:

$$
\begin{aligned}
C(X;Y;Z) = \ & I(X;Y) - I(X;Y|Z) \\
= \ & I(X;Z) - I(X;Z|Y) \\
= \ & I(Y;Z) - I(Y;Z|X)
\end{aligned}
\tag{33}
$$

A similar analysis reveals that the co-information can alternatively be formalized using the intersection inclusion-exclusion criteria for overlapping sets :

$$
\begin{aligned}
C(X;Y;Z) = \ & H(X) + H(Y) + H(Z) \\
& - H(X,Y) - H(X,Z) - H(Y,Z) \\
& + H(X,Y,Z)
\end{aligned}
\tag{34}
$$

This sets up a general pattern for sets of $n$-interacting variables: the entropies of successively larger combinations of variables are added and subtracted based on the parity of the number of variables in the set:

$$C(\mathbf{X}) = -\sum_{\xi \subseteq \mathbf{X}} -1^{(|\mathbf{X}|-|\xi|)} H(\xi) \tag{35}$$

Where $\xi$ indicates all possible subsets of $\mathbf{X}$. Unfortunately, unlike the total and dual total correlations, which easily accommodate large numbers of variables without changing the interpretation, the co-information can be hard to interpret for more than 3 elements. A significant issue is that the co-information can be negative, which has historically limited it's applicability outside of theoretical contexts. [36] showed that, for three variables, the sign of co-information can be understood as indicating whether the statistical dependencies within the dataset are dominated by redundancy or synergy, although this pattern does not necessarily generalize (compare with the O-information, introduced in Sec. VI).

Despite it's practical limitations, however, the co-information has served as a useful jumping off point for research into how mutual information may be further decomposed into redundant and synergistic components (for example, see [23, 46, 47], for a discussion of partial information decomposition, see Sec. IV.

## III.  INFORMATION DYNAMICS

Information dynamics deals with the question of how complex systems "compute" (note the scare quotes), and sits at the intersection of dynamical systems analysis, information theory, computer science, and information theory.

To begin with, what does it mean for a system to "compute?" This inevitably raises the complex, philosophical question "what is computation", which we will sidestep entirely by saying that when a complex system computes, what it is really doing is deciding its immediate next state. This happens at every scale of the system: individual elements can be in multiple states and decide their immediate next state based on the interactions between their own internal dynamics, as well as all the other elements that our single element is causally connected to. The joint-state of the whole system makes up the macro-state of the whole which "emerges" from the interactions of all it's constituent elements [48].

Information dynamics are typically broken down into three components:

1. Information storage [49], which characterizes how the past of a single element constrains it's immediate future,

2. Information transfer [50], which characterizes how the past of a source element constrains the immediate future of a target element,

3. Information modification [51, 52], which describes how multiple streams of information are combined to produce an output that is not trivially reduable to any input.

These three measures can be seen as *very roughly* analogous to the fundamental processes at work in a digital computer: information can be actively stored in memory and retrieved, information can be relayed from one processor to another, and information can be combined in the form of a "computation," to produce some output. When modelling a complex system, understanding the interplay between information storage, transmission, and modification can help elucidate how the behaviour of individual components constrains the evolution of the system as a whole.

For this section, we will be focusing on the information dynamics of a single target (which is the most well-developed area of the field at present). Recently, the notion of information dynamics has been generalized to account for redundant and synergistic interactions between

29

multiple variables: we will briefly discuss this in Sec. **??**, and interested readers can consult [53].

## A. Conditional Entropy Rate

The simplest information dynamics measure is the conditional entropy rate [54, 55], which measures how much uncertainty about the next state of $X$ remains after the past $k$ states are accounted for:

$$H_\mu(X) = \lim_{k\to\infty} H(X_t|X_{-k:t-1}) \tag{36}$$

As $k \to \infty$, the entropy rate gives a measure of unpredictability inherent in $X$ which can never be resolved simply by observing $X$ itself, no matter how long $X$ is watched. For continuous processes, $H_\mu$ is approximated by a number of measures, including sample entropy and approximate entropy [55], and for discrete processes, can be approximated using the Lempel-Ziv complexity [56].

As with all entropy measures, it is possible to calculate the local conditional entropy rate associated with a particular configuration of $X$:

$$h_\mu(x) = \lim_{k\to\infty} h(x_t|x_{-k:t-1}) \tag{37}$$

This tells us how surprising any particular realization $x$ is, given that it is preceded by the particular sequence $x_{-k:t-1}$. The local conditional entropy rate has been used in thermodynamic interpretations of the transfer entropy measure [57].

The local conditional entropy rate can also been used to "whiten" an autocorrelated signal [58]. Using this technique, a time series is transformed by replacing the value at every frame with the local conditional entropy rate of that frame, resulting in a time series with no autocorrelation.

## B. Information Storage

The conditional entropy quantifies the "left over" uncertainty intrinsic to $X$ after the past has been accounted for. Its dual measure is the active information storage [49, 55], which quantifies how much information the *past* of a variable discloses about it's immediate

next state. In this context, "information storage" can be thought of as a measure of how dependent an element of a system is on it's own past. If every next state is chosen at random, then $H_\mu(X) = H(X)$ and $A(X) = 0$. Formally, active information storage is given by:

$$A(X) = \lim_{k \to \infty} I(X_{-k:t-1}; X_t) \tag{38}$$

where $k$ indicates the number of discrete past states to take into account. Following the usual formalism, $X_t$ refers to the state of $X$ at time $t$, while $X_{t-1}^{-k}$ indicates a vector of states, stretching from time $t-1$ all the way back to time $t-1-k$.

The active information storage and entropy rate are complementary: the uncertainty about the next state of $X$ ($H(X)$) can be decomposed into $A(X)$ and $H_\mu(X)$ [49]:

$$H(X) = A(X) + H_\mu(X) \tag{39}$$

Intuitively, this shows us that some of our uncertainty about the next state of $X$ could be resolved by looking at $X$'s past, and a residual part that cannot be resolved. This is a key insight when attempting to model complex systems. For processes with a naturally high entropy rate, simply observing the past behaviour will not help predict the future (and consequently, modelling is hard). Active information storage has been applied in a wide range of contexts, from understanding distributed processing in neural networks [55], predictive coding models [59], eye and gaze tracking [60], to understanding the dynamics of football matches [61].

### 1. Local Active Information Storage

The local active information storage describes how knowledge of the specific past states of $X$ informs on our uncertainty that $X$ will adopt the particular next state $x$ [48, 49].

$$a(x_t) = i(x_{-k:t-1}; x_t) \tag{40}$$

An example of local active information storage in biology might be a neuron: after firing an action potential, neurons typically go through a refractory period as the membrane potential is reset and for that duration, no subsequent firings can occur. In this context, knowing that our target neuron fired within the last $k$ milliseconds drastically reduces our

uncertainty about what the next state of the neuron will be: it probably won't fire, because it's within the refractory window [55]. Local active information storage has also been used to understand swarming behaviours of crustaceans such as crabs [62, 63], and synchronization processes [64].

This allows a "time-resolved" picture of how a single variable remembers and computes with it's own past. When $a(x) > 0$, the system is behaving "as we would expect", in that past states are reducing our uncertainty about future states in accordance with the overall statistics of the system. In contrast, when $a(x) < 0$, we are being "misinformed" and the system appears to be transiently "breaking" from it's own past to do something unexpected.

## C. Information Transfer

The second major type of information dynamics used to study complex systems is that of *information transfer*. While active information storage examines a single variable's behaviour through time, information transfer describes how the behaviour of one (or more) elements in a complex system informs the behaviour of another. The most well-established measure of information transfer is the transfer entropy [65] (for a superb and extremely thorough discussion of transfer entropy in all its glory, see [66]).

For two variables, $X$ and $Y$, the transfer entropy $TE(X \to Y)$ measures the amount of information knowing the *past of X* gives us about the *next state of Y*, in the form of a conditional mutual information:

$$TE(X \to Y) = I(X_{-k:t-1}; Y_t | Y_{-l:t-1}) \tag{41}$$

It is well worth the effort to spend some time unpacking this equation, as it contains a number of important, but subtle interpretations. $TE(X \to Y)$ can be thought of as *the amount of information we get about the next state of Y by knowing the past states of X above and beyond what we would get simply knowing the past state of Y*. Phrased differently: how much information does $X$'s past provide about $Y$'s future that is not also provided about $Y$'s past?

Transfer entropy is probably the most widely used information dynamic in the study of complex systems, having been used in studies of biological networks of neurons, [67? –70], cryptocurrencies [71–74], precious metal economics [75], the interaction between social media

sentiment and commodity prices [74], the directed relationship between atmospheric carbon dioxide levels and global warming [76], climate models [23, 77, 78] and more (for further citations, see [66]). Transfer entropy's nonparametric nature makes it particularly suited to analysis of complex systems where non-linearities are known to play a significant role in the dynamics.

Transfer entropy is a non-linear generalization of the well-known Granger Causality [79, 80]. Here, "causality" refers to the statistical, rather than epistemic, definition, and care should be taken to avoid confusing transfer entropy, which is a measure of effective connectivity, with true causality [80–82]. The question of how to extract causal relationships from data is a rich, and rapidly growing field known as *causal inference* (for an excellent primer, see [83]), which is (regrettably) beyond the scope of this introduction.

### 1. Local Transfer Entropy

As with local mutual information and local active information storage, local transfer entropy gives a high degree of spatial and/or temporal resolution to the analysis of interacting, complex processes [50].

$$te(x_t \rightarrow y_t) = i(x_{-k:t-1}; y_t | y_{-l:t-1}) \tag{42}$$

For example, imagine we were interested in the relationship between two companies $(C_1, C_2)$ traded on the stock market. If we did the full transfer entropy, we might find that $TE(C_1 \rightarrow C2) > 0$, however, it is important to remember that this is the *expected information transfer* over the whole dataset and doesn't tell us much about the variability day-to-day. If, instead, we computed the local transfer entropy at every day, we might find that, for the vast majority of days, $te(c_{1t} \rightarrow c_{2t}) \approx 0$, and it is only on particular days that the predictive relationship holds. Armed with this information, we could do a more detailed investigation, knowing that $C_1$ is predictive of $C_2$ only at particular moments in time (which may be days that $C_1$ announces a new product, or reports their quarterly earnings, etc).

While not nearly as frequently used as the general transfer entropy, local transfer entropy has been used to understand the dynamics of cellular automata [50, 84], neural dynamics measured with electrophysiology [85], and animal swarms [86].

As with any mutual information measure, it is possible to condition the bivariate transfer entropy from $X$ to $Y$ on a larger set of variables to account for common driver effects, redundancies, and synergies in the system [66]. The transfer entropy from $X$ to $Y$, conditioned on $Z$ is given by:

$$TE(X \to Y|Z) = I(X_{-k:t-1}; Y_t | Y_{-l:t-1}, Z_{-r:t-1}) \tag{43}$$

The effect of conditioning on $Z$'s past gives us the amount of information $X$'s past gives us about $Y$'s future *above and beyond that which is provided by $Y$'s own past and $Z$'s past.* As with conditional mutual information, conditional transfer entropy is essential when assessing information dynamics in complex systems, which by definition have large numbers of interacting elements. For two variables $X$ and $Y$ belonging to a set of elements $\mathcal{V}$, the *complete transfer entropy* between $X$ and $Y$ ($TE(X \to Y|\mathcal{V})$) is given by conditioning $TE(X \to Y)$ on the past of every other element in $\mathcal{V}$ simultaneously. The conditional transfer entropy plays a key role in the discovery of effective connectivity networks from time-series data (see Sec. V B and [87, 88]). Conditional transfer entropy can be localized in the same way as the standard bivariate transfer entropy.

### D. Information Modification

Of the three major elements of computation, information modification has resisted attempts to develop a satisfactory formalism [48]. The intuition behind information modification is that the interaction of multiple streams of information causes some non-trivial "change" in the nature of information being transmitted, although exactly what the nature of that change is can be hard to pin down.

An early attempt at formalizing information modification was proposed in [51] called the seperable information. A strictly local measure, seperable information fails as a formal definition of information modification and acts instead as a heuristic based on the interaction between active information storage and bivariate transfer entropy. Subsequently, [52] proposed that the synergistic information between a set of sources onto a target could serve as a more robust, formal definition of information modification (for a discussion of synergistic

information, see Sec. IV). Subsequent work in this vein has used synergistic information as a measure of "non-trivial" computation in neural systems [70, 89–94], where "computation" is thought to relate closely to the idea of information modification (the output of a process being, in some sense, an irreducible function of both inputs).

### E. Excess Entropy

The final measure we will describe here is the excess entropy as, in some sense, it subsumes all of the prior dynamics. The active information storage only tells us how much information from the past is used in computing the immediate next state of our variable. However, not all the information present in the past is used at the immediate next state: some may only become relevant in the distant future, and the total amount can be calculated as the mutual information between the entire past and the entire future [95]:

$$E(X) = \lim_{k \to \infty, j \to \infty} I(X_{-k:t-1}; X_{t:l}) \tag{44}$$

When considering the entire lifetime of the variable, this provides something like a measure of the total amount of predictable structure present in the variable. This is an extremely powerful tool with which to characterize a complex system: given the Bayesian interpretation of mutual information, we can intuitively understand the excess entropy as the maximum theoretical limit on our ability to predict the future dynamics of a system given observations of its past [54]. The local predictive information has not be explored much, to the best of our knowledge, although it can easily be calculated as a local mutual information.

In the even that we are dealing with a multivariate system $\mathbf{X} = \{X^1, \ldots, X^N\}$, then the excess entropy incorporates all of the previously mentioned dynamics (the information storage, the information transfer, and the information modification), as it counts *all* of the predictive power that the entire past discloses about the entire future. In Sec. **??**, we will discuss how the excess entropy can be decomposed to account for the full space of dynamic dependencies between variables.

## IV.  PARTIAL INFORMATION DECOMPOSITION (PID)

We have previously made reference to the idea that information can be synergistically or redundantly shared between multiple sets of elements of a complex system, and that this can complicate our attempts to build models based on empirical data. When discussing conditional mutual information II C 2, we saw how conditioning on multiple variables can reveal when information is present in higher-order configurations: in the case of redundancy, conditioning will reduce the observed quantity of information, while in the case of synergy, conditioning will increase observed quantity of information. Similarly, when discussing multivariate transfer entropy (Sec. III C 2) we saw how conditioning on multiple possible parent elements can reveal effective connections that would be invisible when considering only bivariate links in the system.

While conditioning can help account for redundant and synergistic information when building models, conditioning itself doesn't necessarily give us a detailed picture of how information is distributed over all the elements of interest. A set of source elements can provide *both* redundant and synergistic information about a target element simultaneously, and to get a comprehensive breakdown of how information is distributed requires mathematical machinery beyond what is provided by classical Shannon information theory.

### A.  PID Intuition

Given a set of source elements, all of whom synapse onto a single target element, a natural question is understanding how information about the target element is distributed over different combinations of source elements. To develop the intuition in more detail, let's use the simplest possible "complex system": two parents $X_1$ and $X_2$, both of which synapse onto a child element $Y$ (see Fig. **??**).

As described above, we know that all the information both parents provide about $Y$ is aggregated by $I(X_1, X_2; Y)$. This gives us no insights into how that information is distributed over $X_1$ and $X_2$. For example, what information about $Y$ could be learned by observing either $X_1$ or $X_2$ alone? What about the information that is only disclosed by $X_1$? To build a complete portrait of how information is distributed, we would like to break the joint mutual information down into all possible ways information can be shared:
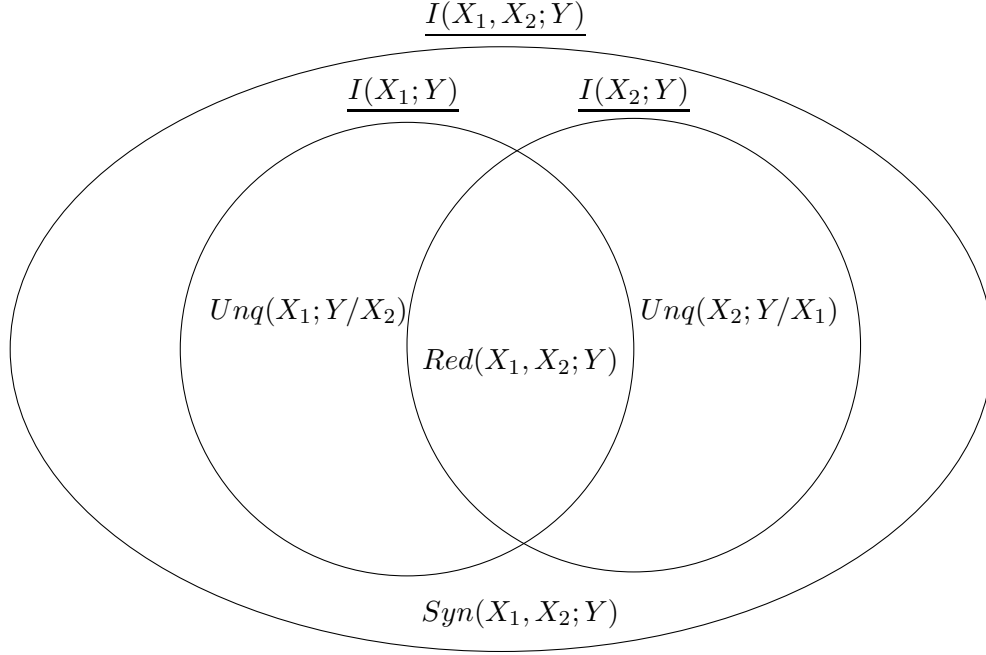
FIG. 3. A Venn Diagram showing how the various components of partial information (redundant, unique, and synergistic) are related to the joint and marginal mutual information terms for two source variables $X_1$ and $X_2$, and a target variable $Y$. The two circles correspond to the mutual information between each source and the target, while the large ellipse gives the joint mutual information between both sources and the target. These diagrams highlight the difference between the marginal mutual information and the unique information: notice that the marginal mutual informations overlap, each one counting the redundant (shared) information towards it's own marginal mutual information. We can also see that $I(X_1, X_2; Y) > I(X_1; Y) \cup I(X_2; Y)$: the difference is the synergistic information which cannot be resolved to either marginal mutual information. Finally, the Venn Diagram highlights how the partial information terms relate to mutual information terms: for example: $Syn(X_1, X_2; Y) = I(X_1, X_2; Y) - I(X_1; Y) - I(X_2; Y) + Red(X_1, X_2; Y)$. We have to add a redundancy term back in because it is "double counted" when subtracting off the marginal mutual informations.

$$I(X_1, X_2; Y) = Red(X_1, X_2; Y)$$
$$+Unq(X_1; Y/X_2)$$
$$+Unq(X_2; Y/X_1)$$
$$+Syn(X_1, X_2; Y) \tag{45}$$

Where $Red(X_1, X_2; Y)$ corresponds to the information about $Y$ that is redundantly present in both $X_1$ and $X_2$ (i.e. could be learned by observing *either* $X_1$ OR $X_2$.). $Unq(X_1; Y/X_2)$ is information about $Y$ that is uniquely present in $X_1$ (i.e. information that could *only* be learned by observing $X_1$), and likequise for $Unq(X_2; Y/X_1)$. Finally, $Syn(X_1, X_2; Y)$ is that information about $Y$ that is synergistically present in the joint states of $X_1$ and $X_2$ (and could not be learned by examining either $X_1$ or $X_2$ individually). A moments reflection should convince the reader that this enumerates all possible combinations.

Furthermore, given such a breakdown, we can decompos each of the marginal mutual informations as well:

$$I(X_1; Y) = Red(X_1, X_2; Y) + Unq(X_1; Y/X_2)$$
$$I(X_2; Y) = Red(X_1, X_2; Y) + Unq(X_2; Y/X_1) \tag{46}$$

The result is a system of three equations (each equal to $I(X_1, X_2; Y)$, $I(X_1; Y)$ and $I(X_2; Y)$ respectively, see Eqs. 45 and 46) and four unknowns ($Red(X_1, X_2; Y)$, $Unq(X_1; Y/X_2)$, $Unq(X_2; Y/X_1)$, and $Syn(X_1, X_2; Y)$). This is an undetermined system: if we were able to calculate any of the four unknowns, we could compute the rest using simple algebra. A number of functions have been proposed targeting the redundancy (see [96] for a review, also Sec. IV B), the unique information (such as [97, 98]), and the synergy (such as [99]).

In general, however, for $N > 2$, it is not the case that the relationship between mutual information terms and partial information "atoms" is so tightly constrained. For $N = 3$, there are seven known mutual information terms but eighteen unique partial information atoms. To solve the general care involves developing some mathematical machinery beyond that which is provided by classical Shannon information.

## B. Redundancy Lattices & Functions

Suppose we have a system of $N$ predictor variables, all informing about the state of a single target $Y$. We would like decompose the joint mutual information $I(X_1, \ldots, K_N; Y)$ down into all relevant combinations of sources. Doing this requites two things: the first is an operational definition of what it means for information to be "redundant" to two or more sources (which we will denote as $I_\cap$). The second is an understanding of the *structure* of multivariate information (which is given by the redundancy lattice described below) and Formally, given some (as yet undefined) set of all meaningful sources of information ($\mathfrak{A}$), we want a decomposition such that:

$$I(X_1, \ldots, X_N) = \sum_{\boldsymbol{\alpha} \in \mathfrak{A}} I_\partial \boldsymbol{\alpha}; Y) \tag{47}$$

Where $I_\partial$ quantifies the information about $Y$ that could only be learned by observing $\boldsymbol{\alpha}$ and no simpler combination of sources.

### 1. The Redundancy Lattice

Here, we will provide a sketch of the derivation of the redundancy lattice first proposed by Williams & Beer [36], and explored further by Gutknecht et al., [10]. Our goal is not a mathematically rigorous derivation, but rather, to provide a basic intuition about where the lattice comes from and what it means. For the interested mathematician, see the above citations.

When decomposing $I(X_1, \ldots, X_N; Y)$, the first step is to enumerate the set of all possible *sources* of information. For example, information can be disclosed by any $X_i$, but it can also be disclosed by any pair $\{X_i, X_j\}$, every triple, and so on. The set $\mathbf{S}$ of all possible (potentially multivriate) sources is given by the power set of the set of $\{X_i\}$: $\mathbf{S} = \mathcal{P}_0(\{X_1, \ldots, X_N\})$, where the 0 subscript indicates that the empty set is being excluded from the power set.

The seminal insight of Williams and Beer was to realize that the problem of decomposing the multivariate information was equivalent to the problem of understanding how information was redundantly distributed over every element of $\mathbf{S}$. This requires computing another power set: this time the power set of $\mathbf{S}$, to get all possible combinations by which elements

39
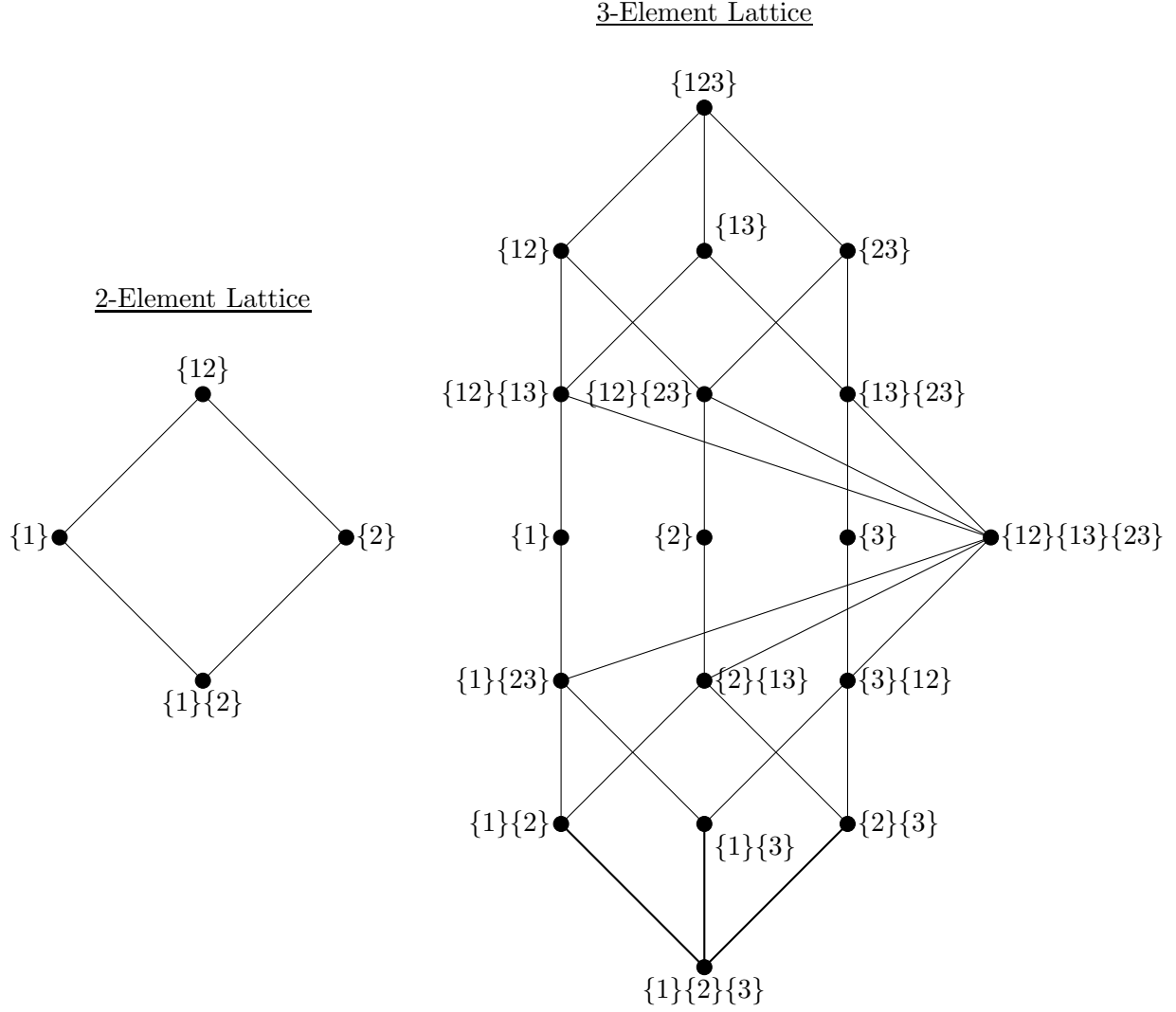
FIG. 4. Examples of redundancy lattices for the two simplest possible systems. **Left:** The redundancy lattice for a set of two sources $X_1$ and $X_2$ synapsing onto a single target. This is a simplified visualization of the Venn Diagram seen in Fig. IV A: {1}{2} corresponds to the information redundantly shared between $X_1$ and $X_2$, while {12} corresponds to the synergistic information and the single elements {1} and {2} indicate the unique information in each element. **Right:** The redundancy lattice for three sources synapsing onto a single target. The three-element lattice makes it clear that, as the number of sources grows, the clean distinctions between "redundancy", "unique information" and "synergy" break down as more complex combinations of sources contribute information about the target. The top and bottom of the lattice can be thought of as "purely synergistic" and "purely redundant" respectively, however between the two extremes, the "PI-atoms" can be thought of as information that is redundantly shared over higher-order combinations of sources: for example {1}{23} gives the partial information that is redundantly present in both $X_1$ and the joint states of $X_2$ and $X_3$ together.

of $\mathbf{S}$ may redundantly share information:

$$\mathfrak{A} = \{\boldsymbol{\alpha} \in \mathcal{P}_0(\mathbf{S}) : \forall \mathbf{A}_i, \mathbf{A}_j \in, \mathbf{A}_i \not\subset \mathbf{A}_j\} \tag{48}$$

Note that not every element of $\mathcal{P}_0(\mathbf{S})$ is a valid member of $\mathfrak{A}$: only those collections of sources where no collection is a proper subset of any other (since, if $\mathbf{A}_i \subset \mathbf{A}_j$, then the information redundantly shared between them would simply be all the information disclosed by $\mathbf{A}_i$). The size of $\mathfrak{A}$ grows hyper-exponentially with the number of elements being considered. For two predictors there are four distinct atoms. For three predictors there are eighteen, and for four predictors there are 166. For eight predictor, there are $\approx 5.6 \times 10^{22}$ atoms. Consequently, the current applicability of the partial information decomposition framework to large systems is limited.

The set $\mathfrak{A}$ has a partially-ordered structure:

$$\forall \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathfrak{A}, \boldsymbol{\alpha} \preceq \boldsymbol{\beta} \Leftrightarrow \forall \mathbf{B} \in \boldsymbol{\beta}, \exists \mathbf{A} \in \boldsymbol{\alpha} \, \text{s.t.} \, \mathbf{A} \subseteq \mathbf{B} \tag{49}$$

This structure allows us to recursively calculate the value of every $\boldsymbol{\alpha} \in \mathfrak{A}$ via Mobius inversion:

$$I_\partial(\boldsymbol{\alpha}; Y) = I_\cap(\boldsymbol{\alpha}; Y) - \sum_{\boldsymbol{\beta} \prec \boldsymbol{\alpha}} I_\partial(\boldsymbol{\beta}; Y) \tag{50}$$

From this, we can extract our initial desired Eq. 47.

The build intuition, let us revisit our toy example from above. If we have $I(X_1, X_2; Y)$, our first step is to enumerate the set of sources:

$$\mathbf{S} = \big\{\{X_1\}, \{X_2\}, \{X_1, X_2\}\big\} \tag{51}$$

We then need to find the domain of our redundancy function $I_\cap$: the set of all sources in $\mathbf{S}$ such that no source is a subset of any other:

$$\mathfrak{A} = \big\{\{X_1\}\{X_2\}, \{X_1\}, \{X_2\}, \{X_1, X_2\}\big\} \tag{52}$$

The foure elements of $\mathfrak{A}$ correspond to our intuitive ideas of redundant, unique, and synergistic information described above. The partial ordering shows that $\{X_1\}\{X_2\} \prec \{X_1\}$ and $\{X_2\}$, while both $\{X_1\}$ and $\{X_2\} \prec \{X_1, X_2\}$. This defines the redundancy lattice show in Fig. IV B 1.

## 2. The Redundancy Function

We have, so far, neglected to define our redundancy function $I_\cap$, instead assuming that behaves in a manner intuitive with our notion of redundancy. To formalize these intuitions, the field has converged on a standard set of axioms that any proposed redundancy function *must* satisfy to induce the lattice structure on $\mathfrak{A}$:

1. **Symmetry:** The function $I_\cap(\mathbf{A}_1, \ldots, \mathbf{A}_k)$ must be invariant under permutation: given a permutation $\sigma$: $I_\cap(\mathbf{A}_1, \ldots, \mathbf{A}_k; Y) = I_\cap(\sigma(\mathbf{A}_1), \ldots, \sigma(\mathbf{A}_k); Y)$

2. **Self-Redundancy:** For single sources, the redundancy is equal to the mutual information: $I_\cap(\mathbf{A}; Y) = I(\mathbf{A}; Y)$.

3. **Monotonicity:** As more sources are added, the redundancy must decrease: $I_\cap(\mathbf{A}_1, \ldots, \mathbf{A}_k, \mathbf{A}_{k+1}; Y)$ $I_\cap(\mathbf{A}_1, \ldots, \mathbf{A}_k; Y)$, with equality if $\mathbf{A}_k \subset \mathbf{A}_{k+1}$.

Any function that satisfies these axioms will induce the partial information lattice described above.

When proposing the partial information decomposition for the first time, Williams and Beer provided a "natural" redundancy function in the form of the *specific information* [36]:

$$I_\cap^{WB}(\boldsymbol{\alpha}; Y) = \sum_{y \in \mathcal{Y}} P(s) I(\boldsymbol{\alpha}; y) \tag{53}$$

While this function has a number of appealing properties (such as returning exclusively non-negative values for the partial information atoms), it was almost immediately criticized for some counter-intuitive behaviours (see [100]). For example, it was considered anomalous that it returned non-zero values in the following case:

$$I_\cap^{WB}(X_1, X_2; \{X_1, X_2\}) \geq 0 \tag{54}$$

Even when $X_1 \perp X_2$. Intuitively, if $X_1$ and $X_2$ are independent, then the information that either one discloses about their joint state $\{X_1, X_2\}$ should be entirely unique. This is called the two-bit copy problem, and the discovery of the apparent flaw prompted widespread interest in alternative redundancy functions. Unlike the Shannon entropy, where a simple set of intuitive axioms is enough to specify a unique, intuitive function, *many* redundancy

functions have been discovered that are consistent with the desired axioms. For a non-exhaustive list, see [23, 46, 96, 100**?** –107]).

These definitions all generally imply slightly different intuitions around the question "what is redundancy". For example, for Gaussian processes, Barrett showed that a natural redundancy function is based on the minimum mutual information [104]:

$$I_\cap^{MMI}(\boldsymbol{\alpha}, Y) = \min_i I(\mathbf{A}_i; Y).$$ (55)

Another significant limitation of $I_\cap^{WB}$ is that it is not localizable in the way that other information in the way that the mutual information is. This has prompted a search for functions that both satisfy the initial axioms, as well as localizability criteria. One of the first was the pointwise common change in surprisal, proposed by Ince [46] and based off the inclusion-exclusion criteria and the co-information:

$$i_\cap^{ccs}(\boldsymbol{\alpha}; Y) = co_Q(\mathbf{a}_1, \ldots, \mathbf{a}_k, y)$$ (56)

Where $co_Q$ refers to the local co-information computed with respect to a maximum-entropy distribution that preserves pairwise marginals. Another, less computationally intensive proposal from Finn & Lizier [28] was based on the logic of informative and misinformative local mutual information. They decomposed the local mutual information into two components corresponding to informative and misinformative probability mass exclusions [28]:

$$i_\cap^\pm(\boldsymbol{\alpha}; y) = \min_i h(\mathbf{a}_i) - \min_i h(\mathbf{a}_i|y)$$ (57)

They showed that, while $i_\cap^\pm$ was not entirely well-behaved (for example, returning negative partial information atoms), the two components, when decomposed individually, satisfied all the original desiderata. This approach was then expanded by Makkeh et al., [107] to a form that is differentiable:

$$i_\cap^{sx}(\mathbf{a}_1, \ldots, \mathbf{a}_k; y) = \log_2 \frac{P(y) - P(y \cap (\bar{\mathbf{a}}_1 \cap \ldots \cap \bar{\mathbf{a}}_k))}{1 - P(\bar{\mathbf{a}}_1 \cap \ldots \cap \bar{\mathbf{a}}_k)} - \log_2 P(y).$$ (58)

Where $P(\bar{\mathbf{a}})$ indicates the probability of seeing $\mathbf{A}$ in any state other than $\mathbf{a}$.

This is an extremely abbreviated survey of redundancy functions. Other, more complex algorithms have been proposed, including ones based on advanced concepts such as infor-

mation geometry [101], probabilistic graphical models [108], or the Gacs-Korner common information [102]. When attempting to implement a PID analysis, care should be taken when choosing a redundancy function, as different functions can return very different results [96, 109].

## C. Information Modification, Revisited

In Sec. III D we discussed how the third information dynamic, information modification, has been discussed in a synergy-based context. We now have the techniques to make this more explicit: for a single element that receives two inputs from parent elements, the information modification can be equated with the synergy [52]. In the context of a dynamic process, synergy can be thought of as the "novel" information generated from the interaction of two streams of information coming together on a single element [94]. Using the same notation as was used in Sec. III, we might say:

$$
\begin{aligned}
Syn(X, Y \to Z) = \lim_{k \to \infty, j \to \infty} & I(X_{-k:t-1}, Y_{-j:t-1}; Z_t) \qquad (59) \\
& - Red(X_{-k:t-1}, Y_{-j:t-1}; Z_t) \\
& - Unq(X_{-k:t-1}; Z_t / Y_{-j:t-1}) \\
& - Unq(Y_{-j:t-1}; Z_t / X_{-k:t-1})
\end{aligned}
$$

For higher order interactions involving more that two variables interacting with a single target, the question of the correct decomposition can be tricky. For example, given three upstream elements all synapsing onto a single target, what should the information modification be? One natural answer would be $I_\partial(\{X_1, X_2, X_3\}$, which is the triple synergy and the most "irreducible" information in the system. However, there are other combinations of sources that use all three elements. For example $I_\partial(\{X_1, X_2\}\{X_1, X_3\}\{X_2, X_3\}$ is the information redundantly distributed over all pairs of joint states. To what extent does this also count as synergistic information modification? As it stands, there is not a single best answer to this question. It may be the case that different atoms communicate different aspects of information processing in complex systems and the idea of a single notion of "information modification" will prove to be overly simplistic.

The PID can also prompt a reinterpretation of the transfer entropy [110]. Recall that:

$$TE(X \to Y) = I(X_{-k:t-1}; Y_t | Y_{-l:t-1}) \tag{60}$$

Using the PID framework, we can decompose the conditional mutual information into:

$$TE(X \to Y) = Unq(X_{-k:t-1}; Y_t) + Syn(X_{-k:t-1}, Y_{-l:t-1}; Y_t) \tag{61}$$

This shows that the transfer entropy consists of two, non-overlapping informational components. The first is referred to as the *state-independent information transfer*: $Unq(X_{-k:t-1}; Y_t)$ is the information that passes from $X$'s past of $Y$'s future and is independent of the state of $Y$'s own past. In contrast, the *state-dependent information transfer*, $Syn(X_{-k:t-1}, Y_{-l:t-1}; Y_t)$, which is the novel information generated at time $t$ in $Y$ by the interaction of $X$'s past and $Y$'s past.

This complicates transfer entropy's status as a measure of information transfer, as it appears to conflate the true transfer of information with synergistic information modification [111]. As it stands, the significance of this has largely remained under-explored. Daube, Gross, and Ince explored the contribution of synergy to sensory EEG data analysis, finding that the apparent transfer entropy was largely reflecting the synergistic component (with the unique "true flow" component being negative) [58]. However, the particular redundancy function they used has been critiqued for itself conflating redundant and synergistic interactions [36, 112], which adds a layer of complexity to the analysis. This is one area in particular with a lot of potential to support future work.

### D.   Extensions of the Partial Information Decomposition

Since the initial proposal by Williams and Beer, there has been work on generalizing the partial information decomposition in various ways. Here we will discuss two: the partial entropy decomposition [47, 113] (which decomposes the joint entropy in a manner analogous to the PID), and the integrated information decomposition [53] (which extends the PID to multiple targets). Other derivatives that we will not discuss in detail include the synergistic disclosure framework [99] (which takes a synergy-first approach rather than a redundancy first approach), a cooperative-game theoretic approach [106] (which uses an alternative lattice and has no notion of redundancy), and an approach based on secret-key cryptography

and dependency lattices [98, 114, 115]. These approaches are, at present, not as developed or widely applied, although each has considerable promise and is worthy for future exploration.

### 1. Partial Entropy Decomposition

The most straightforward generalization the PID is the partial entropy decomposition (PED). First proposed by Ince [47], instead of decomposing the mutual information that some set of predictors discloses about a target, the PED decomposes the joint entropy of a set of variables without requiring categorizing some as sources and others as targets. The general logic is largely the same as the PID: given a redundant entropy function that satisfies all of the axioms described in Sec. IV B 2, it is possible to define a redundancy lattice that can be solved via Mobius inversion in the same way. In a sense, it is related to an early proposed extension of the PID from Bertschinger et al., [100], who proposed a "strong symmetry" axiom that required the redundancy to be insensitive to permutations of sources and targets together, although in the PED framework, the bivariate redundancy $H_\cap(X_1, X_2)$ is not necessarily equal to the mutual information $I(X_1; X_2)$, as in Bertschinger's proposed framework.

The PED reveals interesting, and sometimes counter-intuitive properties of information. For example, consider the decomposition of $H(X_1, X_2)$. As before, we get the decomposition:

$$H(X_1, X_2) = H_\partial(\{X_1\}\{X_2\}) + H_\partial(\{X_1\}) + H_\partial(\{X_2\}) + H_\partial(\{X_1, X_2\}) \tag{62}$$

which again corresponds to a redundant, two unique, and one synergistic atoms. Furthermore, the marginal entropies can be decomposed:

$$H(X_1) = H_\partial(\{X_1\}\{X_2\}) + H_\partial(\{X_1\}) \tag{63}$$

$$H(X_2) = H_\partial(\{X_1\}\{X_2\}) + H_\partial(\{X_2\}) \tag{64}$$

Based on this decomposition, it is possible to re-write the bivariate mutual information in terms of partial entropy atoms:

$$I(X_1; X_2) = H_\partial(\{X_1\}\{X_2\}) - H_\partial(\{X_1, X_2\}) \tag{65}$$

This formulation reveals that the Shannon mutual information can be thought of as the difference between the redundant and synergistic entropies. The exact interpretation of this is difficult, and hinges on what "redundant" and "synergistic" entropy is taken to mean. Ince argued that this showed that mutual information conflated the "true" redundancy and the synergy and prompted a "corrected" mutual information $I^C(X; Y) = I(X; Y) - H_\partial(\{X, Y\})$, although it remains unclear how to interpret this value.

The structure of even bivariate mutual information becomes even more complex as the context in which is exists expands. For example, consider the case of three variables $X_1$, $X_2$, and $X_3$. Following the logic above, it is possible to decompose the bivariate mutual information between two $(I(X_1, X_2))$ into a set of partial entropy atoms that, crucially, can involve redundant and synergistic interactions with the third variable.

$$
\begin{aligned}
I(X_1; X_2) = {} & H_\partial(\{X_1\}\{X_2\}\{X_3\}) + H_\partial(\{X_1\}\{X_2\}) \\
& - H_\partial(\{X_3\}\{X_1, X_2\}) - H_\partial(\{X_1, X_2\}\{X_1, X_3\}\{X_2, X_3\}) \\
& - H_\partial(\{X_1, X_2\}\{X_1, X_3\}) - H_\partial(\{X_1, X_2\}\{X_2, X_3\}) \\
& - H_\partial(\{X_1, X_2\})
\end{aligned}
\tag{66}
$$

Here, we can see that the bivariate mutual information also accounts for information redundantly copied over all three elements ($H_\partial(\{X_1\}\{X_2\}\{X_3\})$), as well as penalizing any information in the joint state of $X_1$ and $X_2$ combined. This shows that, depending on the context, the bivariate mutual information may not be "specific": a dependency between two variables does not necessarily contain only information shared by only those two, but that non-local, "higher-order" information can be present as well. In this case, higher-order information can be counted multiple times, globally inflating the overall amount of apparent dependency in ths system. Varley et al., proposed link between higher-order partial entropy atoms and higher-order data structures such as hypergraphs [112].

The partial entropy decomposition inherent the same unusual property that it requires deciding on one of a number of possible redundant entropy functions. As of this writing, there are three well-explored ones, all of which are defined for local realizations. The first, from Ince [47], is based on the co-information:

$$
h_\cap^{cs}(\boldsymbol{\alpha}) = co_Q(\mathbf{a}_1, \ldots, \mathbf{a}_k)
\tag{67}
$$

With $co_Q$ once again referring to the local co-information computed over a maximum entropy distribution $Q(\mathbf{x})$. This redundancy function can return negative partial entropy values, which is difficult to interpret, although Ince et al., argue that a negative partial entropy atom is an instance of misinformation, similar to the negative local mutual information and can be excluded [47]. The $h_\cap^{cs}$ measure has been critiqued by Varley et al., as being somewhat recursive [112]: it is possible to apply the PED to the redundancy function itself, which reveals that the co-information conflates redundant and synergistic entropy:

$$h_\cap^{cs}(x_1; x_2; x_3) = h_\partial(\{x_1\}\{x_2\}\{x_3\}) \tag{68}$$
$$- h_\partial(\{x_1\}\{x_2, x_3\}) - h_\partial(\{x_2\}\{x_1, x_3\}) - h_\partial(\{x_3\}\{x_1, x_2\})$$
$$- 2 \times h_\partial(\{x_1, x_2\}\{x_1, x_3\}\{x_2, x_3\})$$
$$- h_\partial(\{x_1, x_2\}\{x_1, x_3\}) - h_\partial(\{x_2, x_3\}\{x_1, x_3\}) - h_\partial(\{x_1, x_2\}\{x_2, x_3\})$$
$$+ h_\partial(\{x_1, x_2, x_3\}) \tag{69}$$

Based on this decomposition, we can see that, despite being based on inclusion-exclusion criteria analogous to the intersection of sets, that the co-information does not return only the expected triple-intersection, but rather, conflates the true redundancy with various synergistic terms.

Finn and Lizier independently derived the same framework [113], where they define the redundant entropy of an ensemble as the minimum entropy of any element:

$$h_\cap^{\min}(\boldsymbol{\alpha}) = \min_i h(\mathbf{a}_i) \tag{70}$$

This function returns provably non-negative partial entropy atoms, and can be applied to discrete and continuous data, although it has been criticised for being difficult to interpret. Finally, Varley et al., [112] proposed a redundancy function inspired by the $i_{sx}$ function [107]:

$$h_\cap^{sx}(\boldsymbol{\alpha}) = -\log \frac{1}{P(\mathbf{a}_1 \cup \ldots \cup \mathbf{a}_k)} \tag{71}$$

Like $h_\cap^{min}$, $h_\cap^{sx}$ returns provably non-negative partial entropy atoms, although it is only well-defined for discrete random variables.

$\{12\} \to \{12\}$

$\{1\} \to \{12\}$ $\quad \{2\} \to \{12\}$ $\qquad \{12\} \to \{1\}$ $\quad \{12\} \to \{2\}$

$\{2\} \to \{1\}$

$\{1\}\{2\} \to \{12\}$ $\;\{1\} \to \{1\}$ $\qquad \{2\} \to \{2\}$ $\;\{12\} \to \{1\}\{2\}$

$\{1\} \to \{2\}$

$\{1\}\{2\} \to \{1\}$ $\;\{1\}\{2\} \to \{2\}$ $\qquad \{1\} \to \{1\}\{2\}$ $\quad \{2\} \to \{1\}\{2\}$

$\{1\}\{2\} \to \{1\}\{2\}$
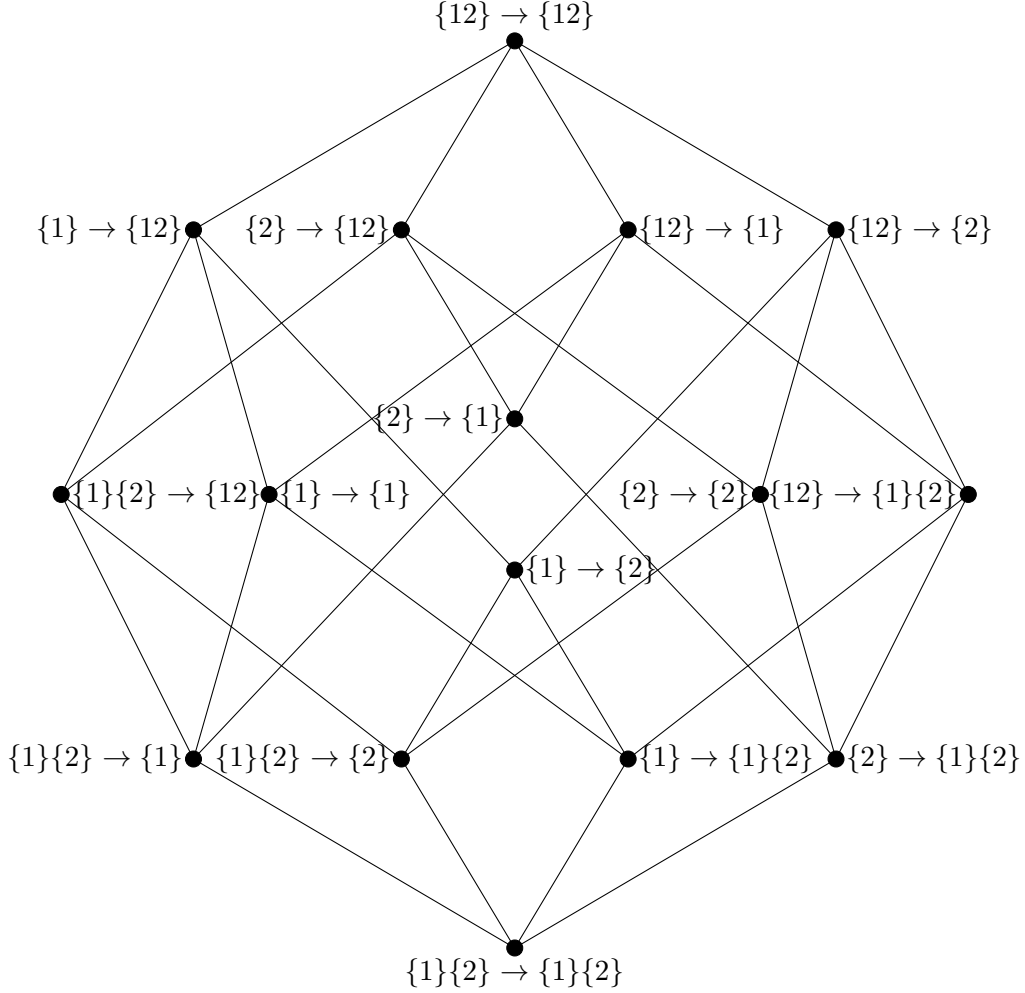
One of the benefits of $h_\cap^{sx}$ is that it can be readily interpreted in the more well-understood context of the PID. Given the identity that: $i(\mathbf{x}; \mathbf{x}) = h(\mathbf{x})$, it can be shown that:

$$h_\cap^{sx}(\mathbf{x}) = i_\cap^{sx}(x_1, \ldots, x_k; \mathbf{x}) \tag{72}$$

For example, $h_\cap^{sx}(\{x_1\}\{x_2\})$ is equivalent to asking "what information about the joint state of $\{x_1, x_2\}$ could be learned by observing either just $x_1$ or just $x_2$?"

Given it's comparative novelty, the PED is much less explored than the PID. While the formal mathematics is well-developed, deep questions of interpretation remain and the application of the framework to real-world complex systems is an area of ongoing research.

### 2. Integrated Information Decomposition

The second generalization of the PID is the *integrated* information decomposition ($\Phi$ID) [53], which expands the PID to multiple sources and multiple targets. Doing so requires proposing a "double redundancy" function that captures some notion of the information that is doubly-redundant across multiple sources and targets. With such a redundancy function, it is possible to derive a double redundancy lattice (see Fig. IV D 2) that can be solved in the usual way via Mobius inversion.

The double redundancy lattice is a *product lattice* $\mathcal{A}^2 = \mathcal{A} \times \mathcal{A}$ (where $\mathcal{A}$ is the single-target redundancy lattice derived above). Each vertex represents an ordered pair $\boldsymbol{\alpha} \to \boldsymbol{\beta}$, with $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{A}$.

The product lattice is a partially ordered set, with:

$$\boldsymbol{\alpha} \to \boldsymbol{\beta} \preceq \boldsymbol{\alpha}' \to \boldsymbol{\beta}' \Leftrightarrow \boldsymbol{\alpha} \preceq \boldsymbol{\alpha}', \boldsymbol{\beta} \preceq \boldsymbol{\beta}' \tag{73}$$

In the $\Phi$ID framework, the double-redundancy lattice is not derived from the properties of redundancy function $I_{\phi\cap}(\boldsymbol{\alpha} \to \boldsymbol{\beta})$ (as is the case in the single-target PID) but rather, it emerges from the product of the "marginal" lattices. This is a non-trivial difference. For the PID, while there has been much contention over the correct definition of "redundancy", essentially everyone agrees that the logic of the lattice is correct (although for an alternative, see [96]) and that the idealized redundancy function "works". Not so for the $\Phi$ID, though.

This does not mean that any redundancy function is valid: there are constraints on what kind of redundancy function will work with the derived lattice. For example, Mediano et al., [53] give a "compatibility constraint" that requires that, if the case of multiple sources and a single target, the $\Phi$ID should revert to a standard PID. Formally, given two variables $\mathbf{X}$, $\mathbf{Y}$ and two double-redundancy atoms $\alpha, \beta \in \mathcal{A}^2$

$$I_{\phi\cap}(\boldsymbol{\alpha} \to \boldsymbol{\beta}) = \begin{cases} I_\cap(\mathbf{X}^{\alpha_1}, \ldots, \mathbf{X}^{\alpha_k}; \mathbf{Y}^{\beta_1}) & \Longleftrightarrow |\mathbf{Y}| = 1 \\ I_\cap(\mathbf{Y}^{\beta_1}, \ldots, \mathbf{Y}^{\beta_j}; \mathbf{X}^{\alpha_1}) & \Longleftrightarrow |\mathbf{X}| = 1 \\ I(\mathbf{X}; \mathbf{Y}) & \Longleftrightarrow |\mathbf{X}| = |\mathbf{Y}| = 1 \end{cases}$$

The compatibility axiom requires that, if one of the variables ($\mathbf{X}$ or $\mathbf{Y}$) is univariate, then the double redundancy function reduces to a classic, single-target redundancy function, and

the $\Phi$ID reduces to the classic PID. This sets up the PID as essentially a "special case" of the $\Phi$ID (a nice property for a proposed generalization).

As it stands, there has been very little empirical work on the integrated information decomposition. Luppi et al., applied it to brain data from healthy individuals [116], as well as those with reduced level of consciousness from either brain injury [117] or anesthesia [118], and found that loss of consciousness was generally associated with a reduction in temporal synergy from the joint state of the past to the joint state of the future. Varley applied the $\Phi$ID framework to spiking neural data [119], finding that the dynamics of temporal synergy and redundancies varied over the course of a neuronal avalanche. At the time of this writing, these are essentially "it."

Given the paucity of work on $\Phi$ID generally, the question of what a natural redundancy function might be has recieved limited attention. Luppi et al., used a minimum-mutual information based double-redundancy function:

$$I_{\phi\cap}^{MMI}(\boldsymbol{\alpha} \to \boldsymbol{\beta}) = \min_{i,j} I(\mathbf{A}_i; \mathbf{B}_j) \tag{74}$$

which can be computed for continuous variables under Gaussian assumptions, but does not readily localize to individual realizations. For the application (fMRI BOLD data), the minimum mutual information is a natural one: it is known to be "correct" in the case of single-target PID for Gaussian variables [104], and is easy to compute.

Varley proposed an alternative measure [119] that is localizable, but only well-defined for discrete data:

$$i_{\phi\cap}^{sx}\tau sx(\boldsymbol{\alpha}; \boldsymbol{\beta}) = \log_2 \frac{P(\mathbf{b}^1 \cup \ldots \cup \mathbf{b}^m)) - P((\mathbf{b}^1 \cup \ldots \cup \mathbf{b}^m) \cap (\bar{\mathbf{a}}^1 \cap \ldots \cap \bar{\mathbf{a}}^k))}{1 - P(\bar{\mathbf{a}}^1 \cap \ldots \cap \bar{\mathbf{a}}^k)} - \log_2 P(\mathbf{b}^1 \cup \ldots \cup \mathbf{b}^m) \tag{75}$$

Varley's measure, built off of the $i_\cap^{sx}$ measure from [107] has the appealing property that it can be decomposed into a sum of partial entropy atoms: $i_{\phi\cap}^{\tau sx}(\boldsymbol{\alpha}; \boldsymbol{\beta}) = h_\cap^{sx}(\boldsymbol{\alpha}) + h_\cap^{sx}(\boldsymbol{\beta}) - h_\cap^{sx}(\boldsymbol{\alpha} \cap \boldsymbol{\beta})$, where $h_\cap^{sx}(\boldsymbol{\alpha} \cap \boldsymbol{\beta})$ satisfies the compatibility axioms and ensures consistency with the product lattice. Since these partial entropy atoms are readily interpretable, this helps ground $i_{\phi\cap}^{\tau sx}(\boldsymbol{\alpha}; \boldsymbol{\beta})$.

*3. Information Dynamics, Revisited*

The $\Phi$ID framework, like the PID framework, is agnostic to the type of data to which it is applied. However (at the time of thsi writing), all published applications of the $\Phi$ID have been to temporal data: pairs of elements that are evolving through time: [116–119]. In this case, the $\Phi$ID can be understood as decomposing the excess entropy (see Eq. 44). The most common approach is lag-1 excess entropy for two variables: $I(X^1_{t-1}, X^2_{t-1}; X^1_t, X^2_t)$. In this case, the various integrated information atoms take on particular interpretations in terms of information dynamics [53]:

**Information storage:** Information in one source at time $t-1$ that *stays* in that source at time $t$. E.g. $\{X_i\} \to \{X_i\}$, $\{X_1\}\{X_2\} \to \{X_1\}\{X_2\}$, and $\{X_1, X_2\} \to \{X_1, X_2\}$.

**Information transfer:** Information in one element at time $t-1$ that moves to another element at time $t$. E.g. $\{X_i\} \to \{X_j\}$.

**Information copying:** Information that is in just one element at time $t$ that is duplicated to be redundantly present in both elements at time $t + 1$. E.g. $\{X_i\} \to \{X_1\}\{X_2\}$.

**Information erasure:** information that is redundantly present in both elements at time $t$ that is "pruned" from one, and left in the other. E.g. $\{X_1\}\{X_2\} \to \{X_i\}$.

**Upward causation:** When information in a "lower order" atom is moved "upward" into a synergistic atom. E.g. $\{X_i\} \to \{X_1, X_2\}$, or $\{X_1\}\{X_2\} \to \{X_1, X_2\}$.

**Downward causation:** When information in a "higher order" atom moves "down" to influence lower-order atoms. E.g. $\{X_1, X_2\} \to \{X_i\}$, or $\{X_1, X_2\} \to \{X_1\}\{X_2\}$. This one is philosophically contentious: the question of whether "downward causation" is a real phenomenon is a topic of active discussion [120, 121].

**Causal decoupling:** A particularly unusual form of information storage that gets it's own category: irreducible information in the "whole" that persists through time: $\{X_1, X_2\} \to \{X_1, X_2\}$. This has also been discussed in the context of "emergence" [53, 122].

In theory, for temporal data, each $\Phi$I atom corresponds to a unique information dynamic. As the number of elements gets larger, however, the intuitive interpretations get harder. For example, for a three-element system, we might propose a dynamic:

**Computation:** When information in the joint state of two sources influences the output of a third. E.g. $\{X_1, X_2\} \rightarrow \{X_3\}$.

This definition of computation is analogous to the one proposed by Timme in the context of the single-target PID [89], and later explored by Sherill [91–94]. Other atoms are more difficult to classify. For a three element system, how might one understand $\{X_1\}\{X_2\}\{X_3\} \rightarrow \{X_1, X_2\}\{X_1, X_3\}\{X_2, X_3\}$? This is the information redundantly copied over all three nodes at time $t$ that is converted into being copied over all pairs of bivariate joint states at time $t + 1$?

Like the PED, the comparative lack of investigation into the $\Phi$ID can be seen as a blessing and a curse. As an under-developed set of tools, off-the-shelf applications are probably pretty few and far between at this point. The flip side of this, however, is that there is much fertile ground to be explored asking fundamental questions and exploring basic applications of these frameworks to complex systems.

## V. NETWORK INFERENCE

One of the foundational methods in complex systems science is the construction of network models from data [123, 124]. It is hard to overstate how fundamental networks are to the modern study of complex systems, appearing in everything from brain science [125] to sociology [126, 127], from ecology [128, 129] to economics [130]. MORE CITATIONS HEREIn applications, networks generally break down into two broad classes: there are physical networks (where the nodes and edge are unambiguous, physical objects in the real world), and statistical networks (where the links between nodes reflects some statistical dependency between them). Physical networks are generally reasonably easy to infer the structure of. For example, the airline network is conceptually simple: nodes may be airports (or cities), and a directed edge exists between two nodes if there is a regular flight from one to another [131]. Another example might be a structural connectivity network map of the brain: using diffusion tensor imaging (DTI) it is possible to get an estimate of the white matter tracts in the brain that physically connect disparate brain regions [132].

In contrast, statistical networks are more abstract: rather than referring directly to some physical structure in the real world (like a white-matter tract, or airline route), the edges in a statistical network refer to some degree of statistical dependency between them. In an information-theoretic context, these dependencies almost always take the form of some kind of mutual information: nodes are usually said to be connected if knowing *something* about the state of one node reduces our uncertainty about the state of another node. There are other "kinds" of statistical network not based on mutual information (for example, dynamic causal modeling build statistical networks using Bayesian model comparison [133, 134]), however for our purposes, we will focus on information-theoretic approaches. Within the category of statistical networks, we will describe two general classes: the "functional" connectivity network, which has no notion of temporal order and produces undirected networks, and the "effective" connectivity network, which does take time into account (if time is a relevant feature) and produces directed networks. Finally, we will discuss how the problem of synergies in data can complicate the very idea of a bivariate network model of a complex system and briefly touch on higher-order models such as hypergraphs and simplicial complexes (although a rigorous mathematical treatment of these topics is beyond the scope of this tutorial).

### A. Functional Connectivity Networks

The simplest form of network construction is that of a "functional connectivity" (FC) network. In an FC network, the edges between nodes are *undirected* and describe some symmetric statistical relationship between the two elements of a complex system. Functional connectivity analysis is generally most well known as a tool in computational neuroscience [125, 135] and an Internet search for "functional connectivity" will return overwhelmingly neuroscience-focused results, however the notion of using time-series statistics to infer a network is much more broadly applicable and has been used when constructing genomic networks [136, 137], financial networks [138, 139], and climate networks [140]. DO I NEED MORE EXAMPLES?

The basic logic of functional connectivity is the equivalence between a correlation matrix and the adjacency matrix of a network. For a system with $N$ features, the correlation matrix is an $N \times N$ matrix that is symmetric about the diagonal (since the correlation is *usually* symmetric in its arguments. If we imagine a network with $N$ nodes, we can draw a link between every pair of nodes, where the weight of edge $i, j$ is equivalent to the corresponding correlation in cell $i, j$. This link between square, symmetric matrices and networks forms the basis of a large field of mathematics known as spectral graph theory (for reference, see [141]), although for our purposes, the isomorphism is sufficient. For a data set with $N$ features (such as $N$ time series, each corresponding to a dynamic element), the basic algorithm for network inference is very simple. In Python, a basic implementation using `numpy` arrays might be:

```
1    N0 = array.shape[0]
2    mi_matrix = np.zeros((N0, N0))
3    for i in range(N0):
4        for j in range(i):
5            mi = mutual_information(array[i], array[j])
6            mi_matrix[i][j] = mi
7            mi_matrix[j][i] = mi
```

The result is a dense matrix, corresponding to a fully-connected network (there is a non-zero edge between every pair of nodes).

Given an efficient estimator of mutual information, this appears to be the end of the line.

However, despite the apparently simplicity, there are two, surprisingly subtle issues that must be accounted for: the question of how to determine if an edge (dependency) is "real", and the inherent bias in the naive estimation of mutual information.

### 1. Bias in Naive Entropy Estimators

The most common way to estimate entropies and mutual information from discrete data is through a naive estimator, where $P(X = x)$ is a function of how many times $x$ appears in a sample, divided by the total number of samples. Due to finite size effects, however, the naive entropy $\hat{H}(X)$ systematically *under-estimates* the true entropy, and by extension, *over-estimates* the true mutual information. Being strictly positive, the naive mutual information estimator is known to be upwardly-biased, reliably *over-estimating* the true mutual information [66]. This means that, unless large amounts of data are included in the mutual information estimates that define the edge weights in the functional connectivity network, the overall information-density of the system is being over-estimated.

Depending on the specific problem being studied, this may not be a problem. Often, as scientists, we don't actually care about the exact value of the mutual information in bits or nats, but rather, the *relative* values of some set of correlations compared to another. For example, if one is interested in comparing brain data recorded from a conscious person versus data from the same individual after being anesthetized (as in [118]), the fact that all the inferred mutual information values are upwardly inflated may not be a critical issue (so long as the upward bias is uniform).

If accurate estimates of the mutual information are critical, there are two possible solutions. The first is analytic corrections. There is a large body of mathematical research dedicated to correcting the inherent bias in naive entropy and mutual information estimators [142], and analytic corrections can typically be implemented with little-to-no computational overhead. The second is to construct a "null" distribution that quantifies the expected null mutual information if there was no correlation between variables. Due to the upward bias of naive mutual information estimators, the expected value of this distribution will be greater than zero, and can be subtracted off from the empirical mutual information to improve the estimate [143, 144].

## 2. *Significance Testing & Thresholding*

The other concern when inferring functional connectivity networks is whether an edge is "real", and deciding on a criteria for whether to include an edge in the network model. The most common way to do this is to compute an analytic or bootstrap a p-value for a null-hypothesis significance test. For Gaussian and discrete variables, it is possible to compute analytic p-values based on Chi-squared distributions [145, 146]. The estimators are infrequently used, although are implemented in the JIDT package [147]. In the absence of an analytic p-value, null hypothesis significance testing can be done by constructing a null distribution (as discussed above and described in detail below). Given some suitably large number of nulls, it is easy to compute an empirical p-value by quantifying the number of null mutual information values greater than or equal to the empirical value.

A final point of discussion is whether a functional connectivity network should be sparsified beyond excluding non-significant edges. This is particularly popular in the human neuroimaging literature, where functional connectivity networks are routinely thresholded at upwards of ninety percent (keeping only the top one to ten percent of edges) [148]. This is often done to justify the use of graph-theoretic measures, which typically are most well-defined on sparse networks. However, recent work has shown that this kind of arbitrary thresholding can produce the appearance of non-trivial topology even in unstructured networks: for example, Cantwell et al., found that thresholding random networks at high values can create plausibly heavy-tailed degree distributions, even when the generating data was Gaussian noise [149]. Other systematic studies of thresholding have also found that it can induce systemic biases in the topology of the resulting network (e.g. [150–152]), and as such, this author recommends *against* sparsifying functional connectivity networks beyond significance testing edges. At the risk of being polemical, radically sparsifying a network to make it receptive to a particular graph-theoretic algorithm is probably indicative that one is putting the cart before the horse: analyses should be constructed to suit the nature of the data, rather than torturing the data to make one's preferred analysis possible. We should not paint with too broad a brush, however, and there may be cases where such choices are appropriate. In general, rather than providing a prescriptive checklist of DO's and DON'Ts, we hope that researchers critically consider the pros and cons of every choice made in a pipeline, rather than simply acting on autopilot or because it was done in a particular way

by a previously highly-cited paper.

### 3. *Null Models for Testing Empirical Mutual Information*

We have made reference to the idea that one can use null distributions to significance test empirical mutual information values, as well as for correcting the inherent bias associated with naive entropy estimators. The question of building strong null models is a surprisingly deep one, and we will not be able to do it justice in this tutorial (for an accessible review, see [153]). In general, null distributions are built by randomizing the existing data in some way, and re-computing the mutual information some large number of times. The question of how exactly to randomize, however, is where a lot of the nuance comes in. There is a fairly simple maxim that can be of practical use here:

*A null model should ideally preserve every possible feature of your data **except** the one that you are testing.*

What does this mean? In many complex systems, patterns or statistics can be consequences of constraints imposed by other features of the system or data (in evolutionary biology, these patterns are called "spandrels" [154, 155]). When generating a null model to significance-test something like the mutual information, we want to ensure that we include all features of the data that might inflate the apparent correlation *other than* the actual statistical dependency of the data.

A commonly-encountered example of this is in the analysis of time series data. It is well-known that autocorrelation in time series data reduces the effective degrees of freedom of the variable and consequently inflates the correlation between two variables [156, 157]. This means that, if the data being studied is time series data, whatever randomization is used to disrupt the dependence between $X$ and $Y$ should preserve the autocorrelation of both variables (for example, using a circular shift null, rather than shuffling the data). By preserving the autocorrelation but randomizing the joint statistics of the two variables, we ensure that the null mutual informations benefit from the same "bump" caused by the autocorrelation that the empirical mutual information did. This is an apples-to-apples comparison, whereas building a null distribution that destroyed autocorrelation in the surrogate data would be

apples-to-oranges.

There are many different kinds of randomization that have been explored in the literature (the `IDTxl` packages conveniently implements a number of options and describes them in the documentation), and choosing the best null for a given dataset is as much an art as it is a technical concern. Depending on exactly what features of the data must be preserved (such as autocorrelation, marginal entropies, etc), the choice of how best to randomize can become quite nuanced. It is not uncommon for the actual analysis of empirical data to take much less time (and be much easier) than the process of designing and generating the correct null model against which to test that empirical data.

## B. Effective Connectivity Networks

The functional connectivity network has no notion of "time" - it describes the instantaneous correlation between elements and would be the same if you were to permute the pair of time series (so long as the joint state was preserved). To account for temporal dynamics requires an alternative approach, one that can infer directed edges. Historically, this was done using the temporal mutual information, quantifying how much knowing the *past* of $X$ reduces our uncertainty about the *future* of $Y$. The temporal mutual information has a significant limitation, however: it will conflate true information transfer with active information storage if the signals are autocorrelated. Consider the case of two time series: $X = \sin(t)$ and $Y = \cos(t)$. Since these signals are period and offset by $\pi/2$, knowing the state of $X$ at time $t$ uniquely specifies the state of $Y$ at time $t + \tau$, which is indicated by the non-zero temporal information. However, the *same* information about $Y_{t+\tau}$ could also be learned by observing $Y$'s own past. There is no *additional* information disclosed by $X$'s past about $Y$'s future that isn't disclosed by $Y$'s own past.

A more natural measure of temporal information flow is the transfer entropy (see Sec. III C for the introduction to transfer entropy). Here we repeat Eq. 41) for convenience:

$$TE(X \to Y) = I(X_{-k:t-1}; Y_t | Y_{-l:t-1}).$$

As described above, the transfer entropy quantifies the information that $X$'s past discloses about $Y$'s future after $Y$'s own past has been taken into account. Importantly, this can include the unique information that flows from past to future, as well as the synergistic

interaction between both pasts ([110, 111], see Sec. IV D 3 for further discussion).

In contrast to the functional connectivity, which is strictly undirected, effective connectivity networks are directed and model how information *flows* from one element of the system to another. The effective network is *not* usually equivalent to the causal network and it is known that multiple different causal structures can give rise to the same effective structure [80–82], but is nevertheless a useful tool that can provide insights beyond a symmetric functional connectivity analysis.

### 1. Optimizing Parameters

Unlike the functional connectivity, which is parameter-free, the transfer entropy requires deciding on how much of the history from the source and target to consider when computing the conditional mutual information. In Eq. 41, these are the two parameters $k$ (for the source past) and $l$ (for the target past). The question of how best to optimize them is a complex topic, and we will not be able to do it justice here. For readers interested in the gory details, we recommend Bossomaier et al.'s *Introduction to Transfer Entropy* [66]. Here, we will briefly review the general concerns (over/underestimating the true value), as well as describing a simple, serial conditioning algorithm (based on [87, 158]) and briefly discussing null-models.

If $k$ is underestimated and not enough of the source's history is taken into account, then the estimate of the effective connectivity will be erroneously low, as information that takes too long to propagate will be excluded (for example, in the case of axon conduction delays [159]). On the other hand, if $k$ is overestimated, then the effective connectivity will be inflated by finite size effects. For a discrete time series that can adopt $N$ states for a given history dimension of $k$, there will be $N^k$ possible unique past states. For large $N$, $k$, or in the case of small data sets, most configurations will only be seen a small number of times, creating the illusion of deterministic transitions and inflating the apparent transfer entropy. Similarly, if the target time series history embedding dimension $l$ is too short, then long-range autocorrelations might be missed, potentially inflating the apparent transfer entropy by incomplete conditioning. On the other hand, if $l$ is too large, the same finite size effects can create illusory determinism in the target's own information storage dynamics.

Picking the correct embedding then, requires balancing a series of trade-offs. Too little

history accounted for will underestimate the true value, while too much history accounted for can compromise it through finite size effects. One possible solution to this problem is a serial conditioning approach (described in detain in [158]). Briefly, the algorithm works by testing whether each additional bin of history produces a statistically significant increase in the transfer entropy when added. For example, if optimizing the target $Y$'s past, one begins with the active information storage with one bin of history: $I(Y_{t-1}; Y_t)$. Then, one bin of history is added in the form of a conditional mutual information: $I(Y_{t-2}; Y_t|Y_{t-1})$, and this value is significance-tested agianst a null distribution. If the increase in the active information storage is significantly *greater* than would be expected by chance, then the increase is likely to be real and not simply a finite-size effect. This process can be repeated with increasing values of $k$ (conditioning on all previous $k-1$ values each time) until either subsequent values are no longer significant, or a pre-defined $k_{\max}$ is reached. Once the optimal value $k^*$ is selected, the same algorithm can be used to optimize the source dimensionality ($l^*$) using the optimal $k^*$. It is also possible to use the same logic to infer a non-uniform embedding, given some alternative stopping criterion (such as a pre-selected $k_{\max}$) [160].

This approach, while powerful, comes with a nontrivial computational cost. For every successive possible value of $k$ and $l$, there must be a null-hypothesis significance test, typically involving the construction of a large distribution of null values. For large systems with many source-target pairs (each requiring a different value of $l$), this can become prohibitively expensive. The sheer number of significance tests also requires stringent correction for false discoveries, necessitating strict corrections.

### 2. Null Models

In Sec. V A 3, we discussed the importance of null models that preserve as many potentially salient features of the data as possible, while randomizing only the relationship of interest. In the case of transfer entropy, this generally necessitates that only the *source* time series is randomized and that the *target* time series remains unchanged. This is based on the logic that the transfer entropy is the information flow that is revealed only when the past of the target is taken into account. A null model then, must preserve the active information storage of the target. A null that involved randomizing both source and target series would be too liberal. Similarly, the randomization of the source time series should preserve the

autocorrelation in the the data, disrupting only the interaction between the variables. A circular-shift null is generally optimal for this purpose.

In some cases, such as in spiking neural recordings, bursty activity or other global signals can artificially inflate the number of significant directed edges (e.g. in [68, 161**?** ]), which can necessitate further processing of the edges using additional tests, such as the "coincidence index" [161] or the "sharpness index" [162]. In general, whatever network comes out of the effective connectivity inference algorithm should be "sanity checked" against known features of the system (i.e. is it implausibly dense?)

### 3.  Redundancy & Synergy in Effective Connectivity Networks

The bivariate transfer entropy, like the bivariate mutual information, has a significant limitation in the context of network inference: it will double-count redundancies while missing higher-order, synergistic relationships. For example, if $Z_t = X_{t-k} \bigoplus Y_{t-l}$, then $TE(X \to Z) = 0$ bit and $TE(Y \to Z) = 0$ bit, while the joint transfer entropy $TE(X, Y \to Z) = 1$ bit. In this case, there is a synergistic interaction between both parents and the target produced by the logical-XOR computation performed by the target. In the case of a system with synergies, a bivariate transfer entropy network is not a "complete" description of the system. The opposite can happen when there is redundant information: if $X$ and $Y$ are both being driven by a common driver $W$ (which is a "grandparent" element to $Z$, then the influence of $W$ can show up multiple times in the interactions between $X$ or $Y$ and $Z$, inflating both transfer entropies and increasing the "apparent" integration of the whole system.

One possible solution to this is with the *global transfer entropy*, which conditions the bivariate transfer entropy on the state of every other element in the system. If $X$ and $Y$ are both elements of $\mathbf{W}$:

$$ TE_G(X \to Y | \mathbf{W}) = I(X_{-k:t-1}; Y_t | Y_{-l:t-1}, \mathbf{W}_{-q:t-1}^{-\{X,Y\}}). $$

Where $\mathbf{W}_{-q:t-1}^{-\{X,Y\}}$ indicates the past state of every element of $\mathbf{W}$ *excluding* $X$ and $Y$. The global transfer entropy is conceptually very powerful, as all synergiestic dependencies that inform on the relationship $X \to Y$ will be illuminated, and redundant information will be pruned. This can make it useful for the study of synergies in complex systems (as in [163]).

However, for large systems, the amount of data required to estimate the multidimensional probability distribution can be prohibitively large.

To address this, [87] propose a multivariate transfer entropy algorithm that uses robust, multilevel statistical tests and a greedy optimization procedure to identify the minimal set of parent elements $\mathbf{W}^*$ that optimizes the conditional transfer entropy $TE(X \to Y | \mathbf{W}^*)$. In doing so, the mutlivarite transfer entropy algorithm builds a bivariate network, but where the weights of the edge between source $X$ and target $Y$ is "informed by" the interaction between every other parent of $Y$ as well. Analysis with simulated Gaussian and mean-field field models has found that the multivariate transfer entropy does a much better job at extracting the ground truth generating model of the simulated data then the bivariate transfer entropy [158, 164]. The multivariate transfer entropy network inference algorithm has been applied to inferring interaction networks between cryptocurrencies [72], biological neuronal networks [70], and human electrophysiological data [165].

It is my recommendation that the multivariate transfer entropy analysis should be used whenever plausible (the only existing implementation is provided by the `IDTxl` package [88]), but it's limitations should also be made clear. While it is less data-intensive than the global transfer entropy, the optimization of $\mathbf{W}^*$, like the optimization of the embedding dimensions and lags, requires considerable computational effort in the form of multiple sequences of large null distributions.

### C. Hypergraphs, Simplicial Complexes, & Other Higher-Order Frameworks

While the multivariate transfer entropy is sensitive to synergies (and prunes redundancies) in a way that the bivariate transfer entropy is not, it could still be argued that networks are simply too restrictive a model for complex systems rich in synergies. For example, consider our toy logical-XOR system (for simplicity, we will exclude the lags from consideration): $Z = X \bigoplus Y$. If we take a conditioning approach, we will find that $I(X; Z | Y) = 1$ bit and and $I(Y; Z | X) = 1$ bit, which will appear as two, bivariate links in the final network. Unfortunately, looking at the network, it is impossible to differentiate between the case where two links represent a *synergistic* interaction between two parents and a target or the case where both parents have independent, specific relationships with the target.

To fully represent the possible, higher-order dependencies fully, it is necessary to leave

the space of strictly bivariate graphs behind and use other frameworks that are more general [166–168]. There are two commonly explored generalizations of the bivariate network: the *hypergraph* and the *simplicial complex*.

Hypergraphs allow edges to be incident on more than two nodes. A standard network is described by two sets: $\mathbf{V} = \{V_1, \ldots, V_k\}$ which defines the nodes of the network, and $\mathbf{E} = \{(V_i, V_j)|i, j \in k \times k\}$, which defines the edges are pairs of nodes. Importantly, every $E_i \in \mathbf{E}$ is a (potentially ordered) pair consisting only of two entries. If we relax the requirement that the elements of $\mathbf{E}$ can only be pairs, but instead tuples of arbitrary length, we have a hypergraph. In the hypergraph, edges can be incident on arbitrary nodes, and can even be subsets of each-other (i.e. it is possible to have $(X_i, X_j, X_k)$ and $(X_i, X_j)$ defining distinct edges with distinct weights. This makes hypergraphs natural frameworks for assessing higher-order synergies and redundancies in data. For example, Varley et al., used hypergraph community detection [169] to look for patterns in sets of three redundantly or synergistically interacting brain regions [112], while Marinazzo et al., used information theory to infer hypergraphical relationships in psychometric data [170]. Finally, Medina-Mardones et al., propose an analytical framework to asses hypergraphical relationships in a computationally tractable way [171].

The other generalization of networks that we will touch on is the simplicial complex. A simplicial complex can be thought of as analogous to a clique, or a fully-connected subgraph. A simplicial complex is made of up simplexes, with are higher-order generalizations of points (0-simplex), lines (1-simplex), triangles (2-simplex), and tetrahedra (3-simplex). A set of simplexes $\mathbf{K}$ is called a simplicial complex if:

1. The face of every simplex in $\mathbf{K}$ is itself a member of $\mathbf{K}$ and,

2. The intersection of any two faces $\mathbf{K}_1, \mathbf{K}_2 \in \mathbf{K}$ is a face of $\mathbf{K}_1$ and $\mathbf{K}_2$.

For example, if we have a set of three elements $X_1$, $X_2$, and $X_3$, and the triangle $(X_1, X_2, X_3)$ is a member of our simplicial complex $\mathbf{K}$, then so too must every edge and point also belong to $\mathbf{X}$ (i.e. $(X_1, X_2)$, $(X_1, X_3)$, etc). This recursive structure makes simplcial complexes tractable with the mathematical machinery of algebraic topology and there is already a very well-developed toolkit for analyzing these structures [166]. Despite this, to the best of our knowledge, at the time of this writing, there have not been any *direct* applications of simplicial complexes to representing higher-order redundancies and synergies, although a

closely related approach comes from Santoro et al., [172], who apply simplicial complexes to the multivariate structure of so-called "edge time series" in human brain data [33]. Given the nature of the relationship between local correlation and local redundancy [112], we conjecture that Santoro et al's method is likely sensitive to higher order redundancies moreso than higher-order synergies, although this remains an area of active research.

Higher-order network representations of complex systems is still in the early stages of development as a field. Many fundamental questions remain, and while this may mean that there aren't as many "off-the-shelf" tools that can be applied (as there are in network science), the opportunities for foundational insights is still very much open.

## VI.   INTEGRATION, SEGREGATION, & COMPLEXITY

A core feature of complex systems is that they incorporate elements of "integration" and "segregation" [39]. Integration refers to a dynamic where all of the elements of the system are interacting and affecting each-other, while segregation refers to a dynamic where subsets of elements are involved in their own processes that are not shared with other elements. As an example, consider the brain: it is known that particular brain areas are involved in some processes and not others (functionality is segregated between different regions), however, at the same time, the brain is integrated enough that all the different local processes can be combined into an integrated, unified whole: a unitary organism with an (apparently) single consciousness. It has been hypothesized that this balance of integration and segregation is key for healthy brain function [125, 173]. Similarly, in economics, successful firms maintain a healthy balance of segregation (each branch working on it's own mission), while all the work is overseen and broadly directed by a centralized executive office. In global politics, the internal dynamics of individual nations are segregated from each-other by national borders, languages, and cultures, while between-nation integration exists in the form of treaties, trade, and historical entanglements.

This mixture of integration and segregation is an inherently multi-scale phenomenon, with different scales often displaying different biases one way or another. Consider a modular network: within each module, there is high integration (which could be assessed using total correlation), but each module may only be sparely connected to other modules, indicating system-wide higher-scale segregation. Depending on how the network is wired together (for instance, having a small-world [174] or scale-free [175] topology), the structure of the overall dynamics being run on the network will change. As mentioned in Sec. I, multi-scale or scale-free structure is one of the hallmarks of complex systems, and is deeply related to the issue of balancing integration and segregation: a system that is fully integrated at the micro-scale cannot support a "higher" scale (a fully connected network has no higher-order communities), nor can a fully segregated system (a set of non-interacting elements barely qualifies as a system). It is a balance of the two that allows non-trivial, multi-scale structure to emerge.

This has been the foundation of several proposed scalar measures of "how complex" a given system is [39, 176]. The notion of "complexity" as reflecting a balance of integration

and segregation recalls the notion discussed in Sec. I that complexity balances predictable order and randomness (an idea that can be formalized in the context of critical phase transitions [177, 178]).

## A. TSE-Complexity

One canonical, information-theoretic attempt to formalize "complexity" as a balance between integration and segregation is given by the Tononi-Sporns-Edelan (TSE) Complexity [39], which iterates over all bipartitions of a given system to quantify how integration and segregation are distributed.

$$TSE(\mathbf{X}) = \sum_{k=1}^{\lfloor N/2 \rfloor} \langle I(\mathbf{X}_j^k; \mathbf{X}_j^{-k}) \rangle \tag{76}$$

Where $\mathbf{X}_j^k$ indicates the $j^{th}$ set of $k$ elements in $\mathbf{X}$ and $\mathbf{X}_j^{-k}$ indicates the complement of $\mathbf{X}_j^k$ (all elements in $\mathbf{X}$ not in $\mathbf{X}_j^k$). Eq. 76 shows us that the TSE complexity of $\mathbf{X}$ can be thought of as the sum of the average mutual informations across all bipartitions of size $k$, and consequently, the measure is high when there is non-trivial information structure observable at every scale. Every element, or combination of elements, on average, discloses some information about the rest of the system, whether we're talking about single units, or large collections.

The TSE complexity can also be formalized in terms of the total correlation of the whole system $\mathbf{X}$ and the expected total correlation of all subsets of $\mathbf{X}$:

$$TSE(\mathbf{X}) = \frac{1}{N} \sum_{k=2}^{N} \left[ \left( \frac{k-1}{N-1} \right) TC(\mathbf{X}) - \langle TC(\mathbf{X}_j^k) \rangle_j \right] \tag{77}$$

This makes it clear that TSE complexity is high when the average integration of all subsets of size $k$ is *less than* what we would expect if integration increased linearly with subset size, *and* the overall integration of the whole system is high [39, 43]. This describes a dynamic where sets of small elements are generally independent from each-other, but nevertheless, the entire system behaves as an integrated whole. The TSE complexity can be used as an objective function when attempting to build systems that strike an optimal balance between integration and segregation [179].

67

Since the TSE-complexity requires brute-forcing all possible bipartitions of the set of elements, it is impossible to fully compute for large systems. Random sampling of subsets can be used to heuristically estimate the complexity [180], although for large systems, even this approach can be computationally intensive. An alternative approach is to consider only the top "layer" of the system: the difference between the total correlation of the whole, and the expected decrease in total correlation when each element is removed [181, 182]:

$$C(\mathbf{X}) = TC(\mathbf{X}) - \frac{TC(\mathbf{X})}{N} - \mathbb{E}[TC(\mathbf{X}_{-i})] \tag{78}$$

While a fairly crude approximation for systems with many scales, the *description complexity* is a tractable heuristic that can be applied to large data sets. The conceit of removing each element once gives it some conceptual similarities to the dual total correlation (Eq. 32). In fact, Varley et al., [183] proved that;

$$DTC(\mathbf{X}) = N \times C(\mathbf{X}) \tag{79}$$

This is consistent with the intuition discussed in Sec. II D 2 that the dual total correlation, unlike the total correlation is highest in a critical zone between highly redundant structure, and total independence.

### B. O-Information & S-Information

The TSE-complexity can be thought of as a measure of how the dependencies that characterize a system are distributed over the elements at very scale. However, it does not give any insight into the *nature* of those dependencies, which could be redundant (information duplicated over many elements) or synergistic (information stored in the joint states of many elements). Rosas et al. showed that highly redundant and highly synergistic systems can both have equivalently high TSE-complexity [43]. As an alternative measure, Rosas et al., proposed a measure they called the O-information:

$$\Omega(\mathbf{X}) = TC(\mathbf{X}) - DTC(\mathbf{X}) \tag{80}$$

The O-information (which was originally named the "enigmatic information" by James and Crutchfield [184]) has the useful property that if $\mathbf{X}$ is dominated by synergistic inter-

actions, then $\Omega(\mathbf{X}) < 0$ bit and conversely, if $\mathbf{X}$ is dominated by redundant interactions, $\Omega(\mathbf{X}) > 0$ bit. It is 0 bit if $\mathbf{X}$ is completely disintegrated (i.e. all $X_i$ are independent), and it is *only* sensitive to beyond-pairwise dependencies (i.e. the bivariate mutual information doesn't contribute towards $+\Omega$ or $-\Omega$). A negative value of $\Omega(\mathbf{X})$ only indicates that $\mathbf{X}$ is *dominated* by synergistic dependencies - redundant dependencies can also exist, which makes it more of a heuristic than the *complete* decompositions provided by the PID and PED. The O-information has generated considerable interest in diverse fields, including neuroscience [183, 185], psychology [170], and music theory [186]. Unlike the PID, which grows hyperexponentially with the number of elements in the system, the O-information scales much more gracefully and can be computed for several hundred interacting elements (assuming sufficient data). The measure has also been extended in a number of directions, including a modified transfer entropy [187], and a time-lagged, frequency-specific version [188].

Varley et al., [183] showed that the O-information can be re-written in terms of sums and differences of just total correlations:

$$\Omega(\mathbf{X}) = (2 - N)TC(\mathbf{X}) + \sum_{i=1}^{N} TC(\mathbf{X}_{-i}) \tag{81}$$

This can help provide some intuition into what synergy means. The left-side term $(2 - N)TC(\mathbf{X})$ is the total integration of the "whole", duplicated $N - 2$ times (and made negative). The right-hand term "adds back" the total correlation associated with the $N$ possible ways that single elements can be removed. If $\sum_{i=1}^{N} TC(\mathbf{X}_{-i}) > (2 - N)TC(\mathbf{X})$, then there must be correlations in the $\mathbf{X}_{-i}$ that are redundant, and so get double counted. Conversely, if $(2 - N)TC(\mathbf{X})$ is greater, then there is information in the structure of the "whole" that can't be found in the parts when all added together.

The O-information can be analyzed locally, in the same way that the bivariate mutual information can be [186]. For a particular configuration $\mathbf{x}$ for $\mathbf{X}$:

$$\omega(\mathbf{x}) = tc(\mathbf{x}) - dtc(\mathbf{x}) \tag{82}$$

Where $tc$ and $dtc$ are the local total correlation and dual total correlation (which can be computed from local entropies). Generally, the sign of $\omega$ is interpreted in the same way as the in the expected case: a negative sign indicates greater synergy while a positive sign

indicates greater redundancy. The local total correlation and local dual total correlation can both be negative, however (unlike their strictly non-negative expected forms). This can create a situation: suppose that $tc(\mathbf{x}) = -3.0$ bit and $dtc(\mathbf{x}) = -8.0$ bit. In this case, $\omega(\mathbf{x}) = +5.0$ bit and we must ask: if $dtc$ is *more misinformative* than the $tc$, is that the same thing as the overall system being "redundancy dominated"? The exploration of the local O-information at this point is very limited and so there remain fundamental questions waiting to be answered.

A closely related measure to the O-information is the *S-information* (again, coined by Rosas et al., [43]). Formally:

$$\Sigma(\mathbf{X}) = TC(\mathbf{X}) + DTC(\mathbf{X}) \tag{83}$$

Originally studied by James and Crutchfield [184], it was initially referred to by the somewhat tongue-in-cheek term "very mutual information", which reflects the fact that it quantifies all of the dependencies between each neuron and the rest of the system:

$$\Sigma(\mathbf{X}) = \sum_{i=1}^{N} I(X_i; \mathbf{X}_{-i}) \tag{84}$$

The S-information is relatively less explored than the O-information, but analysis of discrete and Gaussian random variables has found that it is extremely tightly correlated ($r \approx 1$, [43, 183]). In fact, $\Sigma(\mathbf{X})$ is a much more effective proxy measure for $TSE(\mathbf{X})$ than the initial description complexity/dual total correlation proposed in [181]. Like the O-information, the S-information can be localized:

$$\sigma(\mathbf{x}) = \sum_{i=1}^{N} i(x_i; \mathbf{x}_{-i}) \tag{85}$$
$$= tc(\mathbf{x}) + dtc(\mathbf{x})$$

As before, the local total correlation and local dual total correlation can be negative, although since the relationship is addative, it is conceptually less difficult: informative dependencies contribute to $\sigma(\mathbf{x})$ while misinformative dependencies reduce it. The S-information (both local and average) has potential as a kind of "total statistical density" measure: quantifying how tightly individual nodes are collectively "embedded" in the system under study.

70

## C. Whole-Minus-Sum Complexity & Integrated Information

The above measures are all based on sums and differences of multivariate generalizations of mutual information (total correlation, dual total correlation), which means that they are not dynamic: they operate on static probabilities distributions. A temporal approach (conceptually very similar to some of the information dynamics explored in Sec. III) is with the measure of *integrated information* proposed by Balduzzi and Tononi [189]. Consider a dynamic system comprised of two elements: $\mathbf{X} = \{X_1, X_2\}$. We have previously defined the excess entropy as the mutual information that $\mathbf{X}$'s past discloses about it's own future (Sec. III E), and shown how the excess entropy can be decomposed with multitarget information decomposition (Sec. IV D 2). A simpler heuristic measure is the difference between the excess entropy and the sum of the two "marginal" temporal mutual informations:

$$\Phi(\mathbf{X}) = E(\mathbf{X}) - \sum_{i=1}^{2} I(X_{-k:t}^i, X_{t:j}^i) \tag{86}$$

If $\Phi(\mathbf{X}) > 0$, then there is "integrated" information about the future of the whole system that is only accessible when the past of the whole system is known. In contrast, if $\Phi(\mathbf{X}) < 0$, then the system is dominated by redundant information duplicated over both elements that swamps any synergistic signal [53]. Like the O-information, the sign in this case is a measure of *dominance*, rather than absolute quantity.

Using the $\Phi$ID framework, Mediano et al., were able to propose a modification to $\Phi(\mathbf{X})$, named $\Phi^R$ (pronounced "fire"):

$$\Phi^R(\mathbf{X}) = \Phi^R(\mathbf{X}) - I_\partial(\{1\}\{2\} \rightarrow \{1\}\{2\}) \tag{87}$$

where $I_\partial(\{1\}\{2\} \rightarrow \{1\}\{2\})$ is the double redundancy described by the integrated information decomposition (Sec: IV D 2). $\Phi^R$ is known to be strictly non-negative (unlike the classic *Phi*), and satisfies the initial desire for a measure that describes how much information as in the joint state but not the marginals (for an in-depth discussion, see [53, 122]).

$\Phi$ was originally posed in the context of *integrated information theory* [190], which make very strong claims linking information quantities and phenomonological consciousness. We will remain agnostic on the c-word here, however, a weaker form of integrated information theory has recently emerged as an object of interest in complex systems more generally [191,

192]. Mediano et al., found that $\Phi^R$ generally tracked intuitive ideas of relative "complexity" in complex systems (elementary cellular automata, flocking models, etc) [192] and proposed $\Phi^R$ as a principled, interpretable measure of dynamical complexity. While $\Phi^R$ is only well-defined for the simplest case of a two-element dynamical system, Mediano et al., leverage an old idea first proposed by Tononi and Sporns: the minimum information bipartiton (the MIB) [193]. For a system of many interacting variables, the MIB is the cut that minimizes the information lost by disintegrated the system. For a system of $N$ elements then, which may be too large to do a full $\Phi$ID analysis of, if one can find the MIB, then it is possible to do a bivairate analysis of both partitions (using either $\Phi$ID or $\Phi^R$) to get a heuristic sense to which the system has synergistic dynamics. Finding the minimum information bipartion is easier said than done, however: it is often an intractable problem, and several different heuristics have been proposed to solve it (see Toker and Sommar for review [194]).

Despite these limitations however, we are of the opinion that "weak integrated information theory" holds tremendous promise that has only begun to be realized. It provides an appealing link between the worlds of information theory and computation and dynamical systems approaches on the other. It also is built on strong axiomatic foundations (Shannon information theory, and extensions), which make measures interpretable in ways that other complexity measures may not be. And finally, being purely based on information theory, it is extremely general and can be applied to a variety of domains.

### D.   Do We Even Need Complexity Measures?

The question of a scalar measure of complexity is a naturally interesting one, and, as mentioned, considerable work has gone into attempting to derive the most "natural" one. It is worth taking a step back and, borrowing from the title of [195] ask: "measures of statistical complexity: why?" As mathematical question, the project of finding a scalar measure that captures the nuances of what it means to be complex is an interesting one, but ultimately it is often unclear what has been learned by announcing that "this system has $x$ units of complexity". While there are some cases where relative changes in complexity can be informative (for example, see [196, 197] for two examples), in general it is more useful to begin by defining what particular process or dynamic one is trying to quantify and going from there. Is it synergy/redundancy dominance? Memory? Integration? Predictability?

As complex systems science continues to mature the ability to answer specific questions about specific aspects of system dynamics will become increasingly important. To this end, in the last 10 years, the topic of information dynamics has begun to crystallize as a dominate subfield of information theoretic research.

## VII.   INFORMATION ESTIMATORS

As a branch of pure mathematics, information theory is largely unconcerned with the practical, day-to-day realities of actually *using* information theoretic measures. Pure mathematicians are generally quite happy to live in the land of beautiful abstraction, where measures like entropy, mutual information, and synergy can be formally defined and studied as objects in their own right. For scientists interested in information-theoretic analyses of real world systems, however, there is an immediate issue that presents itself: given a set of data, how does one go about actually estimating the required probabilities, entropies, and related quantities? This is an issue that many aspiring scientists rarely encounter: in most cases, the problems of estimating probabilities are handled "under the hood" of various closed-form statistical tests. When attempting information-theoretic analysis, however, the questions of accurately estimating probabilities is much close to the surface, and care must be taken to ensure that the estimators being used are valid and their various limitations taken into account.

In this section, we will discuss how to estimate entropies and mutual informations for discrete and continuous data. The discrete case is reasonably straightforward (although, as discussed in Sec. VA, bias associated with finite size effects is a concern). The situation becomes more complex for continuous data. We will discuss parametric (Gaussian) estimators for continuous data, as well as briefly touching on non-parametric estimators using K-nearest neighbor graphs. Each one of these has its own strengths and weaknesses, and critical considerations should be made at the outset of an analysis about what approach is best.

## A. Discrete Signals

Historically, the standard practice to infer the relevant probabilities has been to estimate them from the counts of events, so that $\hat{P}(x_i) = n_i/N$, where $n_i$ is the number of occurrences of the $i^{th}$ event and $N$ is the total number of samples. These estimated probabilities can be plugged into Eq. 2 to give the maximum-likelihood estimate of the entropy given the data:

$$\hat{H}(X) = -\sum_{x \in \mathcal{X}} \hat{P}(x) \log(\hat{P}(x)) \tag{88}$$

This is variously known as the *naive entropy estimator*, the *plugin estimator*, as well as the *maximum likelihood estimator*, and is probably the most commonly used method for estimating entropies from data. Despite it's frequency of use, however, the plugin estimator has fundamental limitations: the most significant of which is that it *systematically underestimates the true entropy of the process.* This is known as *bias* ($B$) and is given by:

$$\mathrm{E}[\hat{H}(X)] \leq H(X) \tag{89}$$

The severity of the bias is a function of both the number of states of the system and the number of samples used to construct the probability distributions [143], and a bias-correction term can be introduced for a bias-corrected estimator:

$$\hat{H}_c(X) = \hat{H}(X) + \frac{M-1}{2S} \tag{90}$$

where $M$ is the size of the state-space (eg. $|\mathcal{X}|$) and $S$ is the number of samples. In cases where the number of samples is much larger than the number of states, the bias term becomes negligible, however as the number of states grows (for example when considering the joint entropy of multiple variables), or when the dataset is small and the number of samples limited, bias can quickly overwhelm the plugin estimate, resulting in incorrect estimates unless bias correction is included.

The mutual information suffers from a similar bias, although unlike the entropy, the plugin estimator of mutual information will generally *over-estimate* the true information quantity. Since the mutual information is often calculated as a difference in multiple entropy measures (eg. Eqs. 18), if the bias in MLE estimates of the entropy is greater than the difference, significant errors can be introduced. A number of bias-corrected estimators of

mutual information exist (see: [198]), although the majority of work has been done on estimating the mutual information from continuous signals, although the risk of bias is also of concern even when looking at naturally discrete data.

## B. Continuous Signals

So far, everything that we have discussed assumes that we are analysing data that is naturally discrete: variables can take on one of a finite number of mutually exclusive states (faces of a die, coin flips, etc) that they switch between according to some probability distribution over all states. While this can be very useful when examining naturally digital processes like spiking neurons or a system like a cell that has discrete states (apoptotic, mitotic, etc), the vast majority of empirical data is continuous, from electro-physiological records to climate data, and often does not naturally fall into a set number of obvious "states" that can be assigned to a discrete probability distribution.

This may seem like a significant blow to information theory's utility when analysing the natural world, however considerable work has been done addressing this issue, and several possible solutions are available. They generally fall into two categories: coarse-graining approaches (binning, point-processing, ordinal partition embedding), and density-based approaches (such as the KSG-estimator for mutual information). In general, density-based estimators are considered to be the gold-standard and if possible, should be the first line of attack on a set continuous data. The coarse-graining estimators are presented mostly due to the frequency that these have been used in the past so that the aspiring scientist understands what has been done, as well as providing an opportunity to discuss the limitations inherent to these kinds of analysis.

### 1. Coarse Graining

Coarse graining, or discretizing, is the process by which every observation of a continuous random variable is mapped to one of a finite number of discrete states. It is generally the easiest approach: once a signal is coarse grained, the naive entropy estimators can be used and difficult questions associated with the differential entropy can be avoided. Because of this ease-of-use, coarse graining is far and away the most commonly seen approach to

information-theoretic analysis of continuous signals. Unfortunately, this approach has some fundamental limitations that, when not critically considered at the outset, can seriously compromise the integrity of an analysis.

*Histogram Binning*

The simplest method of coarse graining continuous data is to bin it, constructing probabilities from the histogram of binned values and feeding those values into the naive entropy estimator described above. The apparent simplicity of this method is misleading however, as a number of problems become apparent after looking more closely. The most significant is the trade-off between bias and information loss. We have already seen that the plug-in estimator for Shannon entropy systematically under-estimates the true entropy and, importantly, that the bias gets worse as the number of bins increases. Consequently, we are left attempting a tricky optimization problem: enough bins that we capture the important information in the signal, but not so many that bias renders our numerical estimate worthless. While histogram-based binning approaches have been widely used, at this point they are largely obsolete and should **not** be used unless no other option is available. More advanced approaches can provide the same insights without the inherent biases. In cases where histogram binning must be used, considerable care should be taken to minimize the risk of bias and ensure that the loss of information from the signal is not too great (for example, as in [112]).

*Point Processing*

An alternative to binning the data is creating a "point-process", which maps the continuous signal to a binary time-series by marking "events" at the most extreme points of the signal, usually above some threshold. Point processing assumes that only a subset of moments (typically unusually high-amplitude samples) are significant or informative about the underlying dynamics, and that the majority of the signal is essentially noise. Assuming the assumptions are valid, point-processing is a much more natural method of compressing a complex, continuous signal, as it only records significant events and does not conflate signal and noise in the same way that histogram binning does.

In fMRI data, point-processing has been found to preserve a surprising amount of information, allowing for the reconstruction of higher-order structures such as canonical functional

connectivity networks [34, 199]. When considering a point-process analysis, care must be taken to set the threshold correctly: too low and noisy, low-amplitude fluctuations will confuse the data, while if it is too high, so few moments will be retained that no information will be preserved. In the absence of an *a priori* reason to select a particular threshold, it would be best to run the analysis using several thresholds, to ensure that the results are reasonably robust to small changes in threshold.

An interesting method for choosing a threshold was used in [196, 200, 201]: the authors compared the z-scored empirical distribution of instantaneous amplitudes to the best-fit Gaussian, and set the threshold as the standard deviation at which the empirical data and the best-fit Guassian began to diverge. This ensures that events greater than the threshold are unlikely to have been generated by noise. This assumes that the events of interest are being drawn from some heavier-than Gaussian distribution (e.g. a power-law or lognormal) contaminated by Gaussian noise. Many processes studied in complex systems science are thought to follow heavy-tailed distributions, so this is often not a bad starting place.

As with the histogram-based binning, the point-process throws out a considerable amount of data: a binary process is inherently less able to encode information than a process with more states. If the assumption that all the important information is in these rare "events" is wrong, huge amounts of signal are irretrievably lost. Furthermore, like histogram binning it assumes that every moment in time is something like an independent draw from an underlying distribution: temporally extended fluctuations are ignored, as only the local maxima are retained. While more principled than histogram-based binning, point processes should be treated with similar caution: in fields like neuroscience where point processes are well-explored, it may be a useful technique, but care should be taken to ensure that all assumptions are satisfied.

*Ordinal Partition Embedding*

Both the histogram-based binning and the point-processing transformation assume that every observation of a time-series is a random draw from some underlying distribution, ignoring the possibility that temporally-extended patterns are meaningful as well. An alternative coarse-graining approach is the *permutation embedding* or *ordinal partition embedding*, which transforms continuous data into a sequence based on temporally-extended amplitude patterns [202].

Constructing ordinal partitions requires embedding the time-series $X_t$ in a $d$-dimensional space (with an optional lag of $\tau$), as is commonly done with Taken's embeddings [203]. Each embedded vector $v_t = [x_t, x_{\tau+t}, x_{2\tau+t}...x_{(d-1)\tau+t}]$

As with the point-processing, the ordinal partition embedding as parameters that need to be set: the embedding dimension $d$ and the temporal lag $\tau$. For a discussion of the practical considerations in selecting these parameters, see: [204–206]. If $d$ is too large, then the size of the state-space balloons and under-sampling is a serious concern, while if it is too small, meaningful temporal dependencies can get lost. Similarly, $\tau$ must be optimized to align with the natural temporal dynamics of the system.

### 2. Differential Entropy

In the original proposal by Shannon [15], a generalization of the discrete entropy to continuous random variables was presented in the form of the *differential entropy*:

$$H^{dx}(X) = -\int_x P(x) \log P(x) dx \tag{91}$$

This satisfies the definition that $H^{dx}(X) = \mathbb{E}[-\log P(x)]$, however, it fails to preserve several other properties that make the discrete entropy a useful measure [45]. For instance, it can be negative (as in the case of very tight Gaussian distributions), and it is not invariant under under rescaling due to the additive nature of the logarithm:

$$H^{dx}(\alpha X) = H^{dx}(X) + \log |\alpha| \tag{92}$$

While the differential entropy is not entirely well-behaved, the differential Kullback-Leibler divergence maintains the desire properties of the discrete divergence:

$$D_{KL}^{dx}(P||Q) = \int_x P(x) \log \frac{P(x)}{Q(x)} dx \tag{93}$$

Unlike the differential entropy, the differnetial relative entropy is stictly non-negative (which implies that the differential mutual information is non-negative as well), and preserved under scaling.

If one has a strong reason to believe that continuous data is drawn from a closed-form, parametric distribution, then all of these differential entropies and derivatives can be cal-

culated analytically, and exact estimators exist for a number of distributions. By far, the most common, and well-developed estimators are for univariate and multivariate Gaussian distributions, which are described below.

### 3. Gaussian Estimators

One of the biggest benefits of information theoretic analysis of is that it is sensitive to nonlinear relationships between interacting variables. In complex systems, which are dominated by nonlinear interactions, this can be significant. However, if a researcher knows that their data is already multivariate-normally distributed (or only cares about linear interactions), analytic relationships between information theoretic and classical parametric measures exist. For the marginal and joint entropies:

$$H^{dx}(X) = \ln(\sigma\sqrt{2\pi e}) \tag{94}$$

$$H^{dx}(X_1, X_2...X_n) = \frac{\ln[(2e)^n|\Sigma|]}{2} \tag{95}$$

Where $\sigma^2$ is the standard deviation and $|\Sigma|$ is the determinate of the $n$-dimensional covariance matrix. In general, for a continuous, univariate distribution on the interval $[-\infty, \infty]$ and with fixed mean $\mu$ and variance $\sigma$, the Gaussian distribution is the maximum entropy distribution. For a multivariate distribution with infinite support, fixed mean, and fixed, pairwise covariances, the multivariate Gaussian is the maximum entropy distribution possible [45].

For largely historical reasons, Guassian estimates of information-theoretic quantities are given in nats, rather bits. For bivariate mutual information, the following equality holds:

$$I^{dx}(X;Y) = \frac{-\ln(1-\rho^2)}{2} \tag{96}$$

Where $\rho$ is the Pearson correlation coefficient between $X$ and $Y$. If $\mathbf{X}$ and $\mathbf{Y}$ are multivariate, the generalized Gaussian mutual information is of the form:

$$I^{dx}(\mathbf{X};\mathbf{Y}) = \frac{1}{2}\log\frac{|\Sigma_X||\Sigma_Y|}{|\Sigma_{XY}|} \tag{97}$$

It is also possible to construct Gaussian estimators of local information values. Recall that, for a univariate Guassian:

$$P(x) = \frac{1}{\sigma\sqrt{2\partial}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{98}$$

And for a multivariate Gaussian:

$$P(x_1, x_2...x_n) = \frac{1}{(2\partial)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}e^{\frac{-(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}} \tag{99}$$

Where $\boldsymbol{x}$ is the vector of $x_1, x_2...x_n$ and $\boldsymbol{\mu}$ is the mean vector. We can then substitute these values in any standard information theoretic function (e.g. Eqs. 21) and extract expected or local values. This substitution trick will work for any information theoretic value, allowing more complex estimators, such as the Gaussian transfer entropy, Guassian total correlation, or Gaussian predictive information.

### 4. Density-Based Estimators Estimators

While the parametric Gaussian estimator can be used to efficiently compute information-theoretic measures for continuous signals, it comes with the significant limitation of being blind to nonlinearities in the data. For time-series from complex systems such as financial markets, neural data, or climatic measurements, this can be a severe limitation. To address this, there are non-parametric estimators of information-theoretic measures based on K-nearest neighbours (KNN) inferences of underlying probability manifolds. The field of manifold learning from sparse data is a complex and difficult one and as such the gory details are beyond the scope of this review. Here we will provide the formulae, and the underlying mathematical intuitions behind the derivations can be found in the cited literature.

With any KNN-based analysis, it is crucial to pre-specify the particular distance function to be used and assess it's appropriateness. While in principle any valid metric could be used for the following estimators, in practice, concerns such as the curse of dimensionality often limit what measures will be useful. Here, we are assuming that the reader is planning to use the $L_i nfty$ (Chebyshev) distance, which is robust to high dimensional data and simplifies some of the equations. Readers interested in alternatives such as the $L_1$ (Manhattan distance) or the $L_2$ (Euclidean distance) should review the cited literature.

## Kozachenko-Leonenko Entropy Estimator

The earliest K-nearest neighbours estimator of the entropy of a single or multivariate time-series is given by the Kozachenko-Leonenko estimator [207, 208]. For a potentially multivariate time-series **X**:

$$\hat{H}(\mathbf{X}) = -\psi(k) + \psi(N) + \frac{1}{N} \sum_{i=1}^{N} \log(d_i) \tag{100}$$

where $\psi$ is the Digamma function, $d_i$ is twice the distance from the $i^{th}$ point to it's $k^{th}$ nearest neighbour, and $N$ is the dimensionality of **X**. The Kozachenko-Leonenko estimator can also be used to calculate the local entropy:

$$\hat{h}(\mathbf{x}) = -\psi(k) + \psi(N) + \log(d_i) \tag{101}$$

Since this estimator requires calculating the K-nearest neighbours for every point in the series, the runtime gets large fast. A naive estimator will run in $O(N^2))$ time, although efficient KNN algorithms, such as a KD-Tree can bring the runtime down to $O(N \log(N))$ time.

## Kraskov, Stogbauer, and Grassberge mutual information Estimator

Kraskov, Sogbauer, and Grassberge provided a generalizable nonparametric estimator of mutual information for continuous signals by building off of the Kozachenko-Leonenko KNN-based entropy estimator [198].

There are two different estimators, the first is

$$\hat{I}^1(X;Y) = \psi(K) - \langle \psi(n_x + 1) + \psi(n_y + 1) \rangle + \psi(N) \tag{102}$$

which is more appropriate for smaller samples. Here $n_x$ and $n_y$ correspond to the number of points in the marginal spaces that fall strictly between $\pm$ the maximum distance from the $i^{th}$ point to it's $K^{th}$ nearest neighbour. The second implementation is:

$$\hat{I}^2(X;Y) = \psi(K) - K^{-1} - \langle \psi(n_x) + \psi(n_y) \rangle + \psi(N) \tag{103}$$

and is more appropriate for larger sample sizes. Here $n_x$ and $n_y$ correspond to the number of points in the marginal space that fall between $\pm$ the distance from the $i^{th}$ point to it's

$K^{th}$ nearest neighbour *on their respective axes*. This estimator can also readily generalize to the total correlation by incorporating more dimensions, although it does not generalize to the dual total correlation.

As with the Kozachenko-Leonenko estimator, the KSG estimator can be localized by looking only at the local neighborhood of a particular point:

$$\hat{i}^1(x;y) = \psi(K) - \psi(n_x + 1) + \psi(n_y + 1) + \psi(N) \tag{104}$$

$$\hat{i}^2(x;y) = \psi(K) - K^{-1} - \psi(n_x) + \psi(n_y) + \psi(N) \tag{105}$$

**KSG Conditional mutual information**

As with the Kozachenko-Leonenko entropy estimator and the KSG mutual information estimator, it is not appropriate to simply substitute the various estimators in to established relationships to construct more complex measure: a new estimator of conditional mutual information must be constructed [209–211]. The two bivariate KSG mutual information estimators can be generalized as follows:

$$\hat{I}^1(X;Y|Z) = \psi(K) - \langle \psi(n_z + 1) - \psi(n_{xz} + 1) + \psi(n_{yz} + 1) \rangle + \psi(N) \tag{106}$$

$$\hat{I}^2(X;Y|Z) = \psi(K) - \frac{2}{K} - \langle \psi(n_z) - \psi(n_{xz}) + n_{xz}^{-1} - \psi(n_{yz}) + n_{yz}^{-1} \rangle \tag{107}$$

The local measures can be calculated as described previously.

From the bivariate and conditional mutual informations, non-parametric estimators of almost all the information dynamics measures described above can be constructed, including active information storage, transfer entropy, multivariate transfer entropy, and predictive information. At the time of this writing, there are no non-parametric estimators for partial information decomposition, or certain measures such as the dual total correlation.

## VIII.    PACKAGES FOR INFORMATION THEORY ANALYSIS OF DATA

A large number of packages exist for information theory analysis, far more than we can exhaustively catalogue and review. Many of these are not maintained and guarantees of

quality are impossible. Here, we provide references and brief descriptions of four, well-established toolboxes written and maintained by experts in the field. We have tried to ensure that the variety of programming languages used in modern scientific research are accounted for, including Java, Python, and MATLAB.

### A.    Java Information Dynamics Toollkit (JIDT)

The Java Information Dynamics Toolkit (JIDT) is one of the most versatile of the packages listed here, and also has a shallower learning curve than either IDTxl or DIT [147]. Written in Java, JIDT is a compiled program, and can be called from a variety of programming environments, including MATLAB, Octave, Python, and Java. Being language agnostic makes it much more accessible than other packages listed here. Like IDTxl and the Neuroscience Informaton Toolbox, JIDT is focused primarily on analysing time-series data, and users can read in time-series for quick analysis of a variety of information dynamics phenomena, including transfer entropy (local and global), active information storage (local and global), mutual information, and more.

Unlike all the other listed packages, JIDT also has a point-and-click GUI, allowing users to manually load in datasets, select parameters, and run the analysis without needing to write the script themselves. The package will also provide the script that replicates the given analysis, which can be copy/pasted for reuse in the future. JIDT will provide the scripts in MATLAB, Octave, Python, and Java. This makes JIDT an excellent "starter package" aspiring complex systems scientists who may lack the advanced Python scripting skills required to dive directly into something like RITT, IDTxl or DIT.

Available at: https://github.com/jlizier/jidt

### B.    Information Dynamics Toolkit xl (IDTxl)

IDTxl is a package built on JIDT specifically aimed at using information theory to infer complex effective networks from time-series data in Python [87, 88]. The package includes several different effective network inference methodologies, including multivariate and bivariate transfer entropy, multivariate and bivariate mutual information, as well as tools to analyse active information storage and basic PID. IDTxl gives users a high degree of

control over the inference pipeline, including choosing what information estimator to use (discrete naive, KSG, etc), as well as the number and type of surrogates used during the intensive null-hypothesis significance testing for significant edges. IDTxl's network inference capability integrates with Networkx, a standard Python library for network science analysis [212].

Available at: https://github.com/pwollstadt/IDTxl

## C. Discrete Information Theory Toolbox (DIT)

The Discrete Information Theory Toolbox (DIT) [213] provides Python implementations of a large number of information theoretic functions, including both standards like entropy and mutual information, as well as a suite of more exotic measures, such as the TSE complexity. Unlike JIDT and IDTxl, DIT is not optimized for time-series analysis: the basic object DIT operates on is a multivariate probability distribution, from which the various measures are calculated.

Of the all the packages mentioned here, DIT is the most well-equipped to handle PID analysis: it comes with a large number of redundancy functions built-in, including $I_{min}$ [36], $I_{mmi}$ [100], $I_{ccs}$ [46], and more, as well as a redundant entropy function for PED [47].

Available at: https://github.com/dit/dit

## D. Neuroscience Information Theory Toolbox

Written in MATLAB, the Neuroscience Information Theory toolbox provides functions for those interested in analysing neural spiking data (typically represented as a binary raster) [19]. The toolbox will compute basic Shannon entropies, mutual information, as well as bivariate transfer entropy for building transfer entropy networks. It also implements PID estimators for two sources and a single target using the $I_{min}$ redundancy function [36]. This toolbox has been used extensively for analysis of multielectrode array recordings: see [19, 90, 91]. While other packages presented here are capable of more advanced analyses (multivariate transfer entropy, more complex PID, etc), the Neuroscience Information Theory Toolbox has a shallower learning curve, and the available MATLAB scripts are useful references for efficient implementation of many information theoretic measures.

## IX.   LIMITATIONS OF INFORMATION THEORY

Throughout this review, we have discussed why information theory is a natural *lingua franca* for the field of complex systems, discussed (in some detail) the kinds of insights information theory can give us, and even what packages might be used to do the analysis on empirical data. We should end, however, with a discussion of when, and why, you should **not** use information theory.  Claude Shannon himself bemoaned the fact that, following his introduction of information theory, it became something of a "bandwagon" [18], with researchers racing to apply the framework in ways that he saw as wholly inappropriate.

We have stressed the fact that information theory is largely model-free (certainly compared to standard parametric statistical analyses commonly used in psychology and sociology), however this can be a double-edged sword: without the ability to leverage existing models, information-theoretic analyses generally require considerably more data (sometimes orders of magnitude more) to reliably infer a relationship. In small-N studies, such as those commonly done in psychology, cognitive science, or sociology, hypothesis testing with information theory will likely be more trouble than it is worth (although Gaussian estimators may help alleviate this issue, at the cost of making some constraining assumptions). In addition to big-data requirements, most information theory measures must be calculated by brute force, manually counting the occurrence and co-occurrence rates of different states, often many thousands of times in a row (for significance testing). This can extend the run-times of information theoretic inferences significantly, leaving the researcher in an unfortunate catch-22: having enough data to be confident of a result often means that the runtime will be considerable.

Furthermore, it is difficult to use information theory to build true "causal models" of a process. Measures like active information storage, transfer entropy, or predictive information aggregate the whole set of statistical relationships into single scalar measures without providing much insight into how, or why, certain information is conserved or transmitted. The standard mantra of "correlation does not imply causation" applies just as much to mutual information and transfer entropy as it does to the Pearson product-moment correlation. This is part of the reason that experts in transfer entropy have stressed that effective con-

nectivity is *not* the same thing as causal connectivity [66]. Different causal models can have identical effective connectivity structures, and information theory, at least in it's current formulation, will not provide the means to distinguish between them [80–82].

Finally, it is vitally important to remember what information *is* and *is not*. We talk naturally about information as if it were a substance: quantifying "how much" information available, how it "flows" through systems and how it can be decomposed and separated into distinct "pools," as in PID. This is a case where at least one of the authors feels that our choice of language may be subtlety introducing certain assumptions and biases that are not entirely appropriate. There is, in some corners of complex systems science, a tendency to talk about "information" as if it were a kind of *elan vital*, some kind of aether-like element from which certain mysterious qualities (such as consciousness) might spring. This is, in our opinion, a mistake, and a misunderstanding of what information is and is not. As we discussed in the Introduction, information theory is fundamentally about how to resolve uncertainty in an often-noisy world and intrinsic requires some notion of an observer who can be more or less certain about the state of the observed. It does not make much sense to ask "what is a single neuron's perspective on integration of information", as the information being integrated only in reference to the observer who is modeling the system (this is a oft-discussed topic in the field of autopoesis, for a fascinating reference, see [214]). Two observers (perhaps with different priors, or different apparatuses) might observe the same system and infer different information structures. These models, while distinct from each-other, both accurately describe how observable statistics reduce the specific uncertainties of both modellers.

## X. CONCLUSIONS

In this piece, we have attempted to provide a robust (although far from comprehensive) introduction to the theory and applications of multivariate information theory to the study of complex systems. In doing so, we hope to make information theory a more widely known and commonly used tool in the toolkit of modern science. We feel that this document should not be the *only* exposure to information theory that an interested reader explores: different authors present the same material in different ways, and there is value in viewing things from different perspectives. Textbooks such as Bossomaier et al., [66], Cover and

Thomas [45], and McKay [26] were all consulted during the writing of this review and are themselves invaluable resources. There are also topics that this review does not even touch on, including maximum entropy models, source coding, information bottlenecks, the links between information entropy and physical entropy, and the role of information theory in machine learning. In time, this document may be expanded to include those, as well as whatever new discoveries are made in the coming years. After decades of development, the interest in information theory and the number of teams working on fundamental topics appears to be accelerating. There is undoubtedly much waiting to be revealed.

---

[1] J. Orlowski, "Chasing Coral," (2017).

[2] J. von Braun, in *Health of People, Health of Planet and Our Responsibility: Climate Change, Air Pollution* edited by W. K. Al-Delaimy, V. Ramanathan, and M. Sánchez Sorondo (Springer International Publishing, Cham, 2020) pp. 135–148.

[3] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, Scientific Reports **10**, 16598 (2020), number: 1 Publisher: Nature Publishing Group.

[4] E. Strubell, A. Ganesh, and A. McCallum, in *Proceedings of the 57th Annual Meeting of the Association for* (Association for Computational Linguistics, Florence, Italy, 2019) pp. 3645–3650.

[5] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (CRC Press, 2018) google-Books-ID: 1kpnDwAAQBAJ.

[6] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. van Nes, M. Rietkerk, and G. Sugihara, Nature **461**, 53 (2009).

[7] I. Dobson, B. A. Carreras, V. E. Lynch, and D. E. Newman, Chaos: An Interdisciplinary Journal of Nonlinear Science **17**, 026103 (2007), publisher: American Institute of Physics.

[8] M. E. J. Newman, Contemporary Physics **46**, 323 (2005), arXiv: cond-mat/0412004.

[9] W. R. Ashby, in *Principles of Self Organization: Transactions of the University of Illinois Symposium* (Springer US, Boston, MA, 1991) pp. 521–536.

[10] A. J. Gutknecht, M. Wibral, and A. Makkeh, arXiv:2008.09535 [cs, math, q-bio] (2020), arXiv: 2008.09535.

[11] E. P. Hoel, Entropy **19**, 188 (2017).

[12] F. E. Rosas, P. A. M. Mediano, A. I. Luppi, T. F. Varley, J. T. Lizier, S. Stramaglia, H. J. Jensen, and D. Marinazzo, Nature Physics , 1 (2022), publisher: Nature Publishing Group.

[13] J. P. Crutchfield, Physica D: Nonlinear Phenomena **75**, 11 (1994), number: 1.

[14] W. Weaver, American Scientist **36**, 536 (1948).

[15] C. E. Shannon, Bell System Technical Journal **27**, 379 (1948), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x.

[16] W. R. Ashby, *An introduction to cybernetics* (New York, J. Wiley, 1956).

[17] J. Gleick, *The Information: A History, a Theory, a Flood* (Knopf Doubleday Publishing Group, 2011) google-Books-ID: 617JSFW0D2kC.

[18] C. Shannon, IRE Transactions on Information Theory **2**, 3 (1956), conference Name: IRE Transactions on Information Theory.

[19] N. M. Timme and C. Lapish, eNeuro **5**, ENEURO.0052 (2018).

[20] R. E. Ulanowicz, Computers & Chemistry **25**, 393 (2001).

[21] J. Harte and E. A. Newman, Trends in Ecology & Evolution **29**, 384 (2014).

[22] F. E. Ruiz, P. S. Pérez, and B. I. Bonev, *Information Theory in Computer Vision and Pattern Recognition* (Springer Science & Business Media, 2009) google-Books-ID: yCNOErJ4ENwC.

[23] A. E. Goodwell and P. Kumar, Water Resources Research **53**, 5920 (2017), _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2016WR020216.

[24] C. H. Bennett, Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of M

[25] M. Prokopenko and J. T. Lizier, Scientific Reports **4**, 1 (2014), number: 1 Publisher: Nature Publishing Group.

[26] D. J. C. MacKay and D. J. C. M. Kay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, 2003) google-Books-ID: AKuMj4PN_EMC.

[27] A. Kolchinsky and D. H. Wolpert, Interface Focus **8**, 20180041 (2018), publisher: Royal Society.

[28] C. Finn and J. T. Lizier, Entropy **20**, 826 (2018), number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[29] I. Čmelo, M. Voršilák, and D. Svozil, Journal of Cheminformatics **13**, 3 (2021).

[30] K. W. Church and P. Hanks, Computational Linguistics **16**, 22 (1990).

[31] X. Luo, Z. Liu, M. Shang, J. Lou, and M. Zhou, IEEE Transactions on Network Science and Engineering **8**, 463 (2021), conference Name: IEEE Transactions on Network Science and Engineering.

[32] L. Gueguen, S. Velasco-Forero, and P. Soille, Journal of Mathematical Imaging and Vision **48**, 625 (2014).

[33] J. Faskowitz, F. Z. Esfahlani, Y. Jo, O. Sporns, and R. F. Betzel, Nature Neuroscience **23**, 1644 (2020), number: 12 Publisher: Nature Publishing Group.

[34] O. Sporns, J. Faskowitz, A. S. Teixeira, S. A. Cutts, and R. F. Betzel, Network Neuroscience **5**, 405 (2021).

[35] H. Matsuda, Physical Review E **62**, 3096 (2000), publisher: American Physical Society.

[36] P. L. Williams and R. D. Beer, arXiv:1004.2515 [math-ph, physics:physics, q-bio] (2010), arXiv: 1004.2515.

[37] S. Watanabe, IBM Journal of Research and Development **4**, 66 (1960).

[38] M. Studený and J. Vejnarová, in *Learning in Graphical Models*, NATO ASI Series, edited by M. I. Jordan (Springer Netherlands, Dordrecht, 1998) pp. 261–297.

[39] G. Tononi, O. Sporns, and G. M. Edelman, Proceedings of the National Academy of Sciences **91**, 5033 (1994

[40] S. H. Strogatz, *Sync: How Order Emerges from Chaos In the Universe, Nature, and Daily Life* (Hachette Books, 2012) google-Books-ID: ZQeZAAAAQBAJ.

[41] K. P. O'Keeffe, H. Hong, and S. H. Strogatz, Nature Communications **8**, 1504 (2017), number: 1 Publisher: Nature Publishing Group.

[42] T. S. Han, Inf. Control. (1975), 10.1016/S0019-9958(75)80004-0.

[43] F. Rosas, P. A. M. Mediano, M. Gastpar, and H. J. Jensen, Physical Review E **100**, 032305 (2019), arXiv: 1902.11239.

[44] W. J. McGill, Psychometrika **19**, 97 (1954).

[45] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, 2012) google-Books-ID: VWq5GG6ycxMC.

[46] R. A. A. Ince, Entropy **19**, 318 (2017), number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[47] R. A. A. Ince, arXiv:1702.01591 [cs, math, q-bio, stat] (2017), arXiv: 1702.01591.

[48] J. T. Lizier, *The Local Information Dynamics of Distributed Computation in Complex Systems*, Springer Theses (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013).

[49] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, Information Sciences **208**, 39 (2012).

[50] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, Physical Review E **77**, 026110 (2008), publisher: American Physical Society.

[51] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, Chaos: An Interdisciplinary Journal of Nonlinear Science **20**, 037109 (2010), publisher: American Institute of Physics.

[52] J. T. Lizier, B. Flecker, and P. L. Williams, arXiv:1303.3440 [nlin, physics:physics] (2013), 10.1109/ALIFE.2 arXiv: 1303.3440.

[53] P. A. M. Mediano, F. Rosas, R. L. Carhart-Harris, A. K. Seth, and A. B. Barrett, arXiv:1909.02297 [physics, q-bio] (2019), arXiv: 1909.02297.

[54] J. P. Crutchfield and D. P. Feldman, Chaos: An Interdisciplinary Journal of Nonlinear Science **13**, 25 (2003) publisher: American Institute of Physics.

[55] M. Wibral, J. Lizier, S. Vögler, V. Priesemann, and R. Galuske, Frontiers in Neuroinformatics **8** (2014), 10.3389/fninf.2014.00001, publisher: Frontiers.

[56] J. M. Amigó, J. Szczepański, E. Wajnryb, and M. V. Sanchez-Vives, Neural Computation **16**, 717 (2004), publisher: MIT Press.

[57] M. Prokopenko, J. T. Lizier, and D. C. Price, Entropy **15**, 524 (2013), number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[58] C. Daube, J. Gross, and R. A. A. Ince, arXiv:2201.02461 [q-bio] (2022), arXiv: 2201.02461.

[59] A. Brodski-Guerniero, G.-F. Paasch, P. Wollstadt, I. Özdemir, J. T. Lizier, and M. Wibral, The Journal of Neuroscience **37**, 8273 (2017).

[60] P. Wollstadt, M. Hasenjäger, and C. B. Wiebel-Herboth, Entropy **23**, 167 (2021), number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[61] O. M. Cliff, J. T. Lizier, X. R. Wang, P. Wang, O. Obst, and M. Prokopenko, in RoboCup 2013: Robot World Cup XVII, Vol. 8371, edited by D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, A. Kobsa, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, D. Terzopoulos, D. Tygar, G. Weikum, S. Behnke, M. Veloso, A. Visser, and R. Xiong (Springer Berlin Heidelberg, Berlin, Heidelberg, 2014) pp. 1–12, series Title: Lecture Notes in Computer Science.

[62] X. R. Wang, J. M. Miller, J. T. Lizier, M. Prokopenko, and L. F. Rossi, (2011), 10.13140/2.1.4967.9363, publisher: Unpublished.

[63] T. Tomaru, H. Murakami, T. Niizato, Y. Nishiyama, K. Sonoda, T. Moriyama, and Y.-P. Gunji, Artificial Life and Robotics **21**, 177 (2016).

[64] A. Mutoh and Y.-P. Gunji, AIP Conference Proceedings **1648**, 580014 (2015), publisher: American Institute of Physics.

[65] T. Schreiber, Physical Review Letters **85**, 461 (2000).

[66] T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, *An Introduction to Transfer Entropy: Information Flow in Complex Systems* (Springer, 2016) google-Books-ID: p8eADQAAQBAJ.

[67] N. Timme, S. Ito, M. Myroshnychenko, F.-C. Yeh, E. Hiolski, P. Hottowy, and J. M. Beggs, PLOS ONE **9**, e115764 (2014), publisher: Public Library of Science.

[68] S. Ito, F.-C. Yeh, E. Hiolski, P. Rydygier, D. E. Gunning, P. Hottowy, N. Timme, A. M. Litke, and J. M. Beggs, PLOS ONE **9**, e105324 (2014), number: 8.

[69] S. Nigam, M. Shimono, S. Ito, F.-C. Yeh, N. Timme, M. Myroshnychenko, C. C. Lapish, Z. Tosi, P. Hottowy, W. C. Smith, S. C. Masmanidis, A. M. Litke, O. Sporns, and J. M. Beggs, Journal of Neuroscience **36**, 670 (2016), publisher: Society for Neuroscience Section: Articles.

[70] T. F. Varley, O. Sporns, S. Schaffelhofer, H. Scherberger, and B. Dann, Proceedings of the National Academy of Sciences **120**, e2207677120 (2023), publisher: Proceedings of the National Academy of Sciences.

[71] T. Dimpfl and F. J. Peter, Physica A: Statistical Mechanics and its Applications **516**, 543 (2019).

[72] A. García-Medina and J. B. H. C., Entropy **22**, 760 (2020), number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[73] A. García-Medina and G. G. Farías, PLOS ONE **15**, e0227269 (2020), publisher: Public Library of Science.

[74] Z. Keskin and T. Aste, Royal Society Open Science **7**, 200863 (2020), publisher: Royal Society.

[75] T. Luu Duc Huynh, Resources Policy **66**, 101623 (2020).

[76] A. Stips, D. Macias, C. Coughlan, E. Garcia-Gorriz, and X. S. Liang, Scientific Reports **6**, 21691 (2016).

[77] M. Oh, S. Kim, K. Lim, and S. Y. Kim, Physica A: Statistical Mechanics and its Applications **499**, 233 (201

[78] H. Tongal and B. Sivakumar, Atmospheric Research **255**, 105531 (2021).

[79] L. Barnett, A. B. Barrett, and A. K. Seth, Physical Review Letters **103**, 238701 (2009), publisher: American Physical Society.

[80] F. A. Razak and H. J. Jensen, PLOS ONE **9**, e99462 (2014), publisher: Public Library of Science.

[81] N. Ay and D. Polani, Advances in Complex Systems **11**, 17 (2008), number: 01.

[82] J. T. Lizier and M. Prokopenko, The European Physical Journal B **73**, 605 (2010).

[83] J. Pearl, M. Glymour, and N. P. Jewell, *Causal Inference in Statistics: A Primer* (John Wiley & Sons, 2016).

[84] F. Goetze and P.-Y. Lai, Physical Review E **100**, 012121 (2019), publisher: American Physical Society.

[85] R. Martínez-Cancino, A. Delorme, J. Wagner, K. Kreutz-Delgado, R. C. Sotero, and S. Makeig, Entropy **22**, 1262 (2020), number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[86] E. Crosato, L. Jiang, V. Lecheval, J. T. Lizier, X. R. Wang, P. Tichit, G. Theraulaz, and M. Prokopenko, Swarm Intelligence **12**, 283 (2018).

[87] L. Novelli, P. Wollstadt, P. Mediano, M. Wibral, and J. T. Lizier, Network Neuroscience **3**, 827 (2019).

[88] P. Wollstadt, J. T. Lizier, R. Vicente, C. Finn, M. Martinez-Zarzuela, P. Mediano, L. Novelli, and M. Wibral, Journal of Open Source Software **4**, 1081 (2019).

[89] N. Timme, W. Alford, B. Flecker, and J. M. Beggs, Journal of Computational Neuroscience **36**, 119 (2014).

[90] N. M. Timme, S. Ito, M. Myroshnychenko, S. Nigam, M. Shimono, F.-C. Yeh, P. Hottowy, A. M. Litke, and J. M. Beggs, PLOS Computational Biology **12**, e1004858 (2016), publisher: Public Library of Science.

[91] S. P. Faber, N. M. Timme, J. M. Beggs, and E. L. Newman, Network Neuroscience , 1 (2018).

[92] S. P. Sherrill, N. M. Timme, J. M. Beggs, and E. L. Newman, Network Neuroscience (Cambridge, Mass.) **4**, 678 (2020).

[93] S. P. Sherrill, N. M. Timme, J. M. Beggs, and E. L. Newman, PLOS Computational Biology **17**, e1009196 (2021), publisher: Public Library of Science.

[94] E. L. Newman, T. F. Varley, V. K. Parakkattu, S. P. Sherrill, and J. M. Beggs, Entropy **24**, 930 (2022), number: 7 Publisher: Multidisciplinary Digital Publishing Insti-

tute.

[95] W. Bialek, I. Nemenman, and N. Tishby, Physica A: Statistical Mechanics and its Applications Proc. Int. W

[96] A. Kolchinsky, Entropy **24**, 403 (2022), number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[97] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, Entropy **16**, 2161 (2014), number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[98] R. G. James, J. Emenheiser, and J. P. Crutchfield, Journal of Physics A: Mathematical and Theoretical **52**, 014002 (2018), number: 1 Publisher: IOP Publishing.

[99] F. E. Rosas, P. A. M. Mediano, B. Rassouli, and A. B. Barrett, Journal of Physics A: Mathematical and Theoretical **53**, 485001 (2020), publisher: IOP Publishing.

[100] N. Bertschinger, J. Rauh, E. Olbrich, and J. Jost, arXiv:1210.5902 [cs, math] , 251 (2013), arXiv: 1210.5902.

[101] M. Harder, C. Salge, and D. Polani, Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics **87**,

[102] V. Griffith, E. K. P. Chong, R. G. James, C. J. Ellison, and J. P. Crutchfield, Entropy **16**, 1985 (2014), number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[103] E. Olbrich, N. Bertschinger, and J. Rauh, Entropy **17**, 3501 (2015), number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[104] A. B. Barrett, Physical Review E **91**, 052802 (2015), publisher: American Physical Society.

[105] C. Finn and J. T. Lizier, Entropy **20**, 297 (2018), number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[106] N. Ay, D. Polani, and N. Virgo, arXiv:1910.05979 [cs, math] (2019), arXiv: 1910.05979.

[107] A. Makkeh, A. J. Gutknecht, and M. Wibral, Physical Review E **103**, 032149 (2021), publisher: American Physical Society.

[108] D. Sigtermans, Entropy **22**, 952 (2020), number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[109] J. W. Kay, J. M. Schulz, and W. A. Phillips, Entropy **24**, 1021 (2022), number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[110] P. L. Williams and R. D. Beer, arXiv:1102.1507 [physics] (2011), arXiv: 1102.1507.

[111] R. G. James, N. Barnett, and J. P. Crutchfield, Physical Review Letters **116**, 238701 (2016), number: 23 Publisher: American Physical Society.

[112] T. F. Varley, M. Pope, M. G. Puxeddu, J. Faskowitz, and O. Sporns, "Partial entropy decomposition reveals higher-order structures in human brain activity," (2023), arXiv:2301.05307 [cs, math, q-bio].

[113] C. Finn and J. T. Lizier, Entropy **22**, 216 (2020), number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[114] R. G. James, J. Emenheiser, and J. P. Crutchfield, Entropy **21**, 12 (2019), number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[115] J. W. Kay and R. A. A. Ince, Entropy **20**, 240 (2018), number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[116] A. I. Luppi, P. A. M. Mediano, F. E. Rosas, N. Holland, T. D. Fryer, J. T. O'Brien, J. B. Rowe, D. K. Menon, D. Bor, and E. A. Stamatakis, Nature Neuroscience , 1 (2022), publisher: Nature Publishing Group.

[117] A. I. Luppi, P. A. M. Mediano, F. E. Rosas, J. Allanson, J. D. Pickard, G. B. Williams, M. M. Craig, P. Finoia, A. R. D. Peattie, P. Coppola, D. K. Menon, D. Bor, and E. A. Stamatakis, "Reduced emergent character of neural dynamics in patients with a disrupted connectome," (2022), pages: 2022.06.16.496445 Section: New Results.

[118] A. I. Luppi, P. A. M. Mediano, F. E. Rosas, J. Allanson, J. D. Pickard, R. L. Carhart-Harris, G. B. Williams, M. M. Craig, P. Finoia, A. M. Owen, L. Naci, D. K. Menon, D. Bor, and E. A. Stamatakis, bioRxiv , 2020.11.25.398081 (2020), publisher: Cold Spring Harbor Laboratory Section: New Results.

[119] T. F. Varley, arXiv:2202.12992 [math, q-bio] (2022), https://doi.org/10.48550/arXiv.2202.12992, arXiv: 2202.12992.

[120] M. A. Bedau, Principia **6**, 5 (2002), publisher: Núcleo de Epistemologia E Lógica – Nel, da Universidade Federal de Santa Catarina – Ufsc.

[121] S. Gibb, R. F. Hendry, and T. Lancaster, *The Routledge Handbook of Emergence* (Routledge, 2019) google-Books-ID: ptSNDwAAQBAJ.

[122] F. E. Rosas, P. A. M. Mediano, H. J. Jensen, A. K. Seth, A. B. Barrett, R. L. Carhart-Harris, and D. Bor, arXiv:2004.08220 [nlin, q-bio] (2020), arXiv: 2004.08220.

[123] A.-L. Barabási and M. PÃ3sfai, *Network Science* (Cambridge University Press, 2016) google-Books-ID: iLtGDQAAQBAJ.

[124] F. Menczer, S. Fortunato, and C. A. Davis, *A First Course in Network Science* (Cambridge University Press, 2020) google-Books-ID: KYHCDwAAQBAJ.

[125] O. Sporns, *Networks of the Brain* (MIT Press, 2010).

[126] M. S. Granovetter, American Journal of Sociology **78**, 1360 (1973), publisher: The University of Chicago Press.

[127] G. Kossinets and D. J. Watts, Science **311**, 88 (2006), publisher: American Association for the Advancement of Science.

[128] J. A. Dunne, R. J. Williams, and N. D. Martinez, Marine Ecology Progress Series **273**, 291 (2004).

[129] R. M. Thompson, U. Brose, J. A. Dunne, R. O. Hall, S. Hladyz, R. L. Kitching, N. D. Martinez, H. Rantala, T. N. Romanuk, D. B. Stouffer, and J. M. Tylianakis, Trends in Ecology & Evolution **27**, 689 (2012), publisher: Elsevier.

[130] A. Vié and A. J. Morales, Computational Economics (2020), 10.1007/s10614-020-10021-5.

[131] W.-B. Du, X.-L. Zhou, O. Lordan, Z. Wang, C. Zhao, and Y.-B. Zhu, Transportation Research Part E: Logistics and Transportation Review **89**, 108 (2016).

[132] E. Garyfallidis, M. Brett, B. Amirbekian, A. Rokem, S. Van Der Walt, M. Descoteaux, and I. Nimmo-Smith, Frontiers in Neuroinformatics **8** (2014), 10.3389/fninf.2014.00008.

[133] K. J. Friston, L. Harrison, and W. Penny, NeuroImage **19**, 1273 (2003).

[134] K. Friston, K. H. Preller, C. Mathys, H. Cagnan, J. Heinzle, A. Razi, and P. Zeidman, Neuroimage **199**, 730 (2019).

[135] A. Fornito, A. Zalesky, and E. Bullmore, *Fundamentals of Brain Network Analysis* (Academic Press, 2016).

[136] A. J. Butte and I. S. Kohane, in *Biocomputing 2000* (WORLD SCIENTIFIC, 1999) pp. 418–429.

[137] K.-C. Liang and X. Wang, EURASIP Journal on Bioinformatics and Systems Biology **2008**, 1 (2008).

[138] P. Fiedor, Physical Review E **89**, 052801 (2014), publisher: American Physical Society.

[139] X. Guo, H. Zhang, and T. Tian, PLOS ONE **13**, e0195941 (2018), publisher: Public Library of Science.

[140] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, The European Physical Journal Special Topics **174**, 157 (2009).

[141] A. E. Brouwer and W. H. Haemers, *Spectra of Graphs*, Universitext (Springer, New York, NY, 2012).

[142] T. Schürmann, Journal of Physics A: Mathematical and General **37**, L295 (2004).

[143] G. Miller, Information theory in psychology : Problems and methods (1955), publisher: Free Press.

[144] R. A. Ince, B. L. Giordano, C. Kayser, G. A. Rousselet, J. Gross, and P. G. Schyns, Human Brain Mapping **38**, 1541 (2016).

[145] D. R. Brillinger, Brazilian Journal of Probability and Statistics **18**, 163 (2004), publisher: [Brazilian Statistical Association, Institute of Mathematical Statistics].

[146] P. E. Cheng, J. W. Liou, M. Liou, and J. A. Aston, Journal of Data Science **4**, 387 (2006).

[147] J. T. Lizier, Frontiers in Robotics and AI **1** (2014), 10.3389/frobt.2014.00011, publisher: Frontiers.

[148] M. P. van den Heuvel, S. C. de Lange, A. Zalesky, C. Seguin, B. T. T. Yeo, and R. Schmidt, NeuroImage **152**, 437 (2017).

[149] G. T. Cantwell, Y. Liu, B. F. Maier, A. C. Schwarze, C. A. Serván, J. Snyder, and G. St-Onge, Physical Review E **101**, 062302 (2020), number: 6 Publisher: American Physical Society.

[150] A. J. Schwarz and J. McGonigle, NeuroImage **55**, 1132 (2011).

[151] K. A. Garrison, D. Scheinost, E. S. Finn, X. Shen, and R. T. Constable, NeuroImage **118**, 651 (2015).

[152] T. Adamovich, I. Zakharov, A. Tabueva, and S. Malykh, Scientific Reports **12**, 18659 (2022), number: 1 Publisher: Nature Publishing Group.

[153] F. Váša and B. Mišić, Nature Reviews Neuroscience , 1 (2022), publisher: Nature Publishing Group.

[154] S. J. Gould, R. C. Lewontin, J. Maynard Smith, and R. Holliday, Proceedings of the Royal Society of London. Series B. Biological Sciences **205**, 581 (1997), publisher: Royal Society.

[155] M. Rubinov, Nature Communications **7**, 13812 (2016), bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Complex networks;Network models Subject_term_id:

complex-networks;network-models.

[156] S. Afyouni, S. M. Smith, and T. E. Nichols, NeuroImage **199**, 609 (2019).

[157] O. M. Cliff, L. E. Novelli, B. D. Fulcher, J. M. Shine, and J. T. Lizier, (2020), publisher: ArXiv.

[158] L. Novelli and J. T. Lizier, arXiv:2007.07500 [physics, q-bio] (2020), arXiv: 2007.07500.

[159] J.-D. Lemaréchal, M. Jedynak, L. Trebaul, A. Boyer, F. Tadel, M. Bhattacharjee, P. Deman, V. Tuyisenge, L. Ayoubian, E. Hugues, B. Chanteloup-Forêt, C. Saubat, R. Zouglech, G. C. Reyes Mejia, S. Tourbier, P. Hagmann, C. Adam, C. Barba, F. Bartolomei, T. Blauwblomme, J. Curot, F. Dubeau, S. Francione, M. Garcés, E. Hirsch, E. Landré, S. Liu, L. Maillard, E.-L. Metsähonkala, I. Mindruta, A. Nica, M. Pail, A. M. Petrescu, S. Rheims, R. Rocamora, A. Schulze-Bonhage, W. Szurhaj, D. Taussig, A. Valentin, H. Wang, P. Kahane, N. George, and O. David, Brain **145**, 1653 (2021).

[160] L. Faes, G. Nollo, and A. Porta, Physical Review E **83**, 051112 (2011), publisher: American Physical Society.

[161] S. Ito, M. E. Hansen, R. Heiland, A. Lumsdaine, A. M. Litke, and J. M. Beggs, PLOS ONE **6**, e27431 (2011), publisher: Public Library of Science.

[162] M. Kajiwara, R. Nomura, F. Goetze, M. Kawabata, Y. Isomura, T. Akutsu, and M. Shimono, PLOS Computational Biology **17**, e1008846 (2021), publisher: Public Library of Science.

[163] D. Marinazzo, L. Angelini, M. Pellicoro, and S. Stramaglia, Physical Review E **99**, 040101 (2019), publisher: American Physical Society.

[164] M. Ursino, G. Ricci, and E. Magosso, Frontiers in Computational Neuroscience **14** (2020), 10.3389/fncom.20 publisher: Frontiers.

[165] D. J. Harmah, C. Li, F. Li, Y. Liao, J. Wang, W. M. A. Ayedh, J. C. Bore, D. Yao, W. Dong, and P. Xu, Frontiers in Computational Neuroscience **13**, 85 (2019).

[166] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri, arXiv:2006.01764 [cond-mat, physics:nlin, physics:physics, q-bio] (2020), arXiv: 2006.01764.

[167] L. Torres, A. S. Blevins, D. Bassett, and T. Eliassi-Rad, SIAM Review **63**, 435 (2021), publisher: Society for Industrial and Applied Mathematics.

[168] C. Bick, E. Gross, H. A. Harrington, and M. T. Schaub, "What are higher-order networks?" (2022), arXiv:2104.11329 [nlin, stat].

[169] T. Kumar, S. Vaidyanathan, H. Ananthapadmanabhan, S. Parthasarathy, and B. Ravindran, in *Complex Networks and Their Applications VIII*, Studies in Computational Intelligence, edited by H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha (Springer International Publishing, Cham, 2020) pp. 286–297.

[170] D. Marinazzo, J. Van Roozendaal, F. E. Rosas, M. Stella, R. Comolatti, N. Colenbier, S. Stramaglia, and Y. Rosseel, "An information-theoretic approach to hypergraph psychometrics," (2022), number: arXiv:2205.01035 arXiv:2205.01035 [stat].

[171] A. M. Medina-Mardones, F. Rosas, S. E. Rodríguez, and R. Cofré, Journal of Physics: Complexity (2021), 10.1088/2632-072X/abf231.

[172] A. Santoro, F. Battiston, G. Petri, and E. Amico, Nature Physics , 1 (2023), publisher: Nature Publishing Group.

[173] G. Tononi, BMC Neuroscience **5**, 42 (2004).

[174] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998), number: 6684 Publisher: Nature Publishing Group.

[175] A.-L. Barabási and R. Albert, Science **286**, 509 (1999), publisher: American Association for the Advancement of Science Section: Report.

[176] N. Ay, E. Olbrich, N. Bertschinger, and U. JOST, ECCS'06 : Proceedings of the European Conference on Complex Systems 2006 (2006).

[177] N. M. Timme, N. J. Marshall, N. Bennett, M. Ripp, E. Lautzenhiser, and J. M. Beggs, Frontiers in Physiology **7** (2016), 10.3389/fphys.2016.00425.

[178] N. J. M. Popiel, S. Khajehabdollahi, P. M. Abeyasinghe, F. Riganello, E. S. Nichols, A. M. Owen, and A. Soddu, Entropy **22**, 339 (2020), number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[179] O. Sporns and M. Lungarella, in *Artificial Life X* (MIT Press, 2006).

[180] N. Marshall, N. M. Timme, N. Bennett, M. Ripp, E. Lautzenhiser, and J. M. Beggs, Frontiers in Physiology **7** (2016), 10.3389/fphys.2016.00250.

[181] G. Tononi, G. M. Edelman, and O. Sporns, Trends in Cognitive Sciences **2**, 474 (1998).

[182] O. Sporns, G. Tononi, and G. M. Edelman, Behavioural Brain Research **135**, 69 (2002).

[183] T. F. Varley, M. Pope, J. Faskowitz, and O. Sporns, "Multivariate Information Theory Uncovers Synergistic Subsystems of the Human Cerebral Cortex,"

(2022), number: arXiv:2206.06477 arXiv:2206.06477 [cs, math, q-bio].

[184] R. G. James, C. J. Ellison, and J. P. Crutchfield, Chaos: An Interdisciplinary Journal of Nonlinear Science **21**, 037109 (2011), publisher: American Institute of Physics.

[185] M. Gatica, R. Cofré, P. A. Mediano, F. E. Rosas, P. Orio, I. Diez, S. P. Swinnen, and J. M. Cortes, Brain Connectivity (2021), 10.1089/brain.2020.0982, publisher: Mary Ann Liebert, Inc., publishers.

[186] T. Scagliarini, D. Marinazzo, Y. Guo, S. Stramaglia, and F. E. Rosas, Physical Review Research **4**, 013184 (2022), publisher: American Physical Society.

[187] S. Stramaglia, T. Scagliarini, B. C. Daniels, and D. Marinazzo, Frontiers in Physiology **11** (2021), 10.3389/fphys.2020.595736, publisher: Frontiers.

[188] L. Faes, G. Mijatovic, Y. Antonacci, R. Pernice, C. Barà, L. Sparacino, M. Sammartino, A. Porta, D. Marinazzo, and S. Stramaglia, IEEE Transactions on Signal Processing **70**, 5766 (2022), conference Name: IEEE Transactions on Signal Processing.

[189] D. Balduzzi and G. Tononi, PLoS computational biology **4**, e1000091 (2008).

[190] G. Tononi, The Biological Bulletin **215**, 216 (2008), number: 3.

[191] P. A. M. Mediano, F. E. Rosas, D. Bor, A. K. Seth, and A. B. Barrett, Trends in Cognitive Sciences **26**, 646 (2022), publisher: Elsevier.

[192] P. A. M. Mediano, F. E. Rosas, J. C. Farah, M. Shanahan, D. Bor, and A. B. Barrett, Chaos: An Interdisciplinary Journal of Nonlinear Science **32**, 013115 (2022), publisher: American Institute of Physics.

[193] G. Tononi and O. Sporns, BMC Neuroscience **4**, 31 (2003), number: 1.

[194] D. Toker and F. T. Sommer, PLOS Computational Biology **15**, e1006807 (2019), number: 2.

[195] D. P. Feldman and J. P. Crutchfield, Physics Letters A **238**, 244 (1998), number: 4.

[196] T. F. Varley, O. Sporns, A. Puce, and J. Beggs, PLOS Computational Biology **16**, e1008418 (2020), number: 12 Publisher: Public Library of Science.

[197] T. F. Varley, V. Denny, O. Sporns, and A. Patania, Royal Society Open Science **8**, 201971 (2021), publisher: Royal Society.

[198] A. Kraskov, H. Stoegbauer, and P. Grassberger, Physical Review E **69**, 066138 (2004), arXiv: cond-mat/0305641.

[199] E. Tagliazucchi, P. Balenzuela, D. Fraiman, and D. R. Chialvo, Frontiers in Physiology **3** (2012), 10.3389/fphys.2012.00015.

[200] C. Meisel, E. Olbrich, O. Shriki, and P. Achermann, The Journal of Neuroscience **33**, 17363 (2013).

[201] O. Shriki, J. Alstott, F. Carver, T. Holroyd, R. N. Henson, M. L. Smith, R. Coppola, E. Bullmore, and D. Plenz, The Journal of neuroscience : the official journal of the Society for Neuroscience **33**, 7079 (2013).

[202] C. Bandt and B. Pompe, Physical Review Letters **88**, 174102 (2002).

[203] F. Takens, in *Dynamical Systems and Turbulence, Warwick 1980*, Vol. 898, edited by D. Rand and L.-S. Young (Springer Berlin Heidelberg, Berlin, Heidelberg, 1981) pp. 366–381, series Title: Lecture Notes in Mathematics.

[204] M. Riedl, A. Müller, and N. Wessel, The European Physical Journal Special Topics **222**, 249 (2013).

[205] M. McCullough, M. Small, and H. H.-C. Iu, , 4 (2015).

[206] M. McCullough, M. Small, T. Stemler, and H. H.-C. Iu, Chaos: An Interdisciplinary Journal of Nonlinear Science **25**, 053101 (2015).

[207] L. F. Kozachenko and N. N. Leonenko, Problems of Information Transmission **23**, 9 (1987).

[208] S. Delattre and N. Fournier, Journal of Statistical Planning and Inference **185**, 69 (2017).

[209] S. Frenzel and B. Pompe, Physical Review Letters **99**, 204101 (2007), publisher: American Physical Society.

[210] M. Wibral, R. Vicente, and M. Lindner, in *Directed Information Measures in Neuroscience*, Understanding Complex Systems, edited by M. Wibral, R. Vicente, and J. T. Lizier (Springer, Berlin, Heidelberg, 2014) pp. 3–36.

[211] G. Gómez-Herrero, W. Wu, K. Rutanen, M. C. Soriano, G. Pipa, and R. Vicente, Entropy **17**, 1958 (2015), number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[212] A. Hagberg, D. Schult, and P. Swart (2008).

[213] R. James, C. Ellison, and J. Crutchfield, Journal of Open Source Software **3**, 738 (2018).

[214] F. G. Varela, H. R. Maturana, and R. Uribe, in *Facets of Systems Science*, International Federation for Systems Research International Series on Systems Science and Engineering,

edited by G. J. Klir (Springer US, Boston, MA, 1991) pp. 559–569.

**APPENDIX**

**Basic Probability Theory**

Probability theory is a branch of mathematics that deals with the question "how likely events are to occur?"

The core structure of study in probability theory is a *random variable*, which can be thought of as any kind of thing that can take on various states. In our case, we will restrict ourselves to *discrete* random variables, where there are only a finite number of possible states that the variable can adopt, and it will only adopt one at a time (states are mutually exclusive). Examples of discrete random variables include coin flips (which can be Heads or Tails), die (which can roll one of six faces), or a deck of cards (from which 52 unique cards can be drawn). For a random variable $X$ ("big-ecks"), we denote specific outcomes as $X = x$ or just $x$ ("little-ecks").

The *support set* of the a random variable is the set of all possible outcomes. For a variable $X$, it's support set is typically denoted as $\mathcal{X}$. The cardinality of the support set ($|\mathcal{X}|$) gives the number of unique possible states our variable can take on. If $X$ is a fair, six-sided die, than $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ and $|\mathcal{X}| = 6$. The support set of a variable is sometimes also referred to as the "image" of that variable.

Every element in the support set has an associated numerical value, between 0 and 1, called the *probability*, denoted as $P(X = x)$. The question of "what is probability" is a deep philosophical one, with ongoing battles between Frequentist and Bayesian philosophers of mathematics and we won't dig into the gory details here. For now, we will take a Bayesian approach and say that the probability of a specific outcome is a measure of our belief about how likely that particular outcome is. If we say that there is a 90% chance of rain tomorrow, we are indicating that we are very confident that it will rain tomorrow. Similarly, if we say there is only a 1% chance of snow, we are indicating that we are very confident that it will not snow. A 50% chance of snow suggests that we are not very confident and it could go either between both binary outcomes (snow and no snow).

By definition, the probabilities of every outcome in $\mathcal{X}$ must sum to 1:

$$\sum_{x \in \mathcal{X}} P(X = x) = 1 \tag{108}$$

*Properties of Probabilities*

For two random variables $X$ and $Y$, we can look at the probabilities of specific pairs of events co-occurring. This is the joint probability and is given by $P(X = x, Y = y)$. As with the individual probabiliteis (called "marginal probabilities"), the joint probabilities must all sum to 1:

$$\sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} P(X = x, Y = y) = 1 \tag{109}$$

We can also calculate how the probabilities of specific outcomes of $X$ changes depending on the state of $Y$ using the conditional probability $P(X = x|Y = y)$. Two examples can help make the intuition behind conditional probability clear: first, imagine two fair independent coins. The outcome of the first coin flip was Heads and we're interested in whether than affects the probability that the second coin flip will be heads. Since the coins are fair and independent, the outcome of the first flip has no effect on the second, so $P(X = Heads|Y = Heads) = 1/2$. For a second example, imagine a deck of 52 shuffled playing cards. We draw two cards and are interested in whether either one of them is the Ace of Spades. For the first card, $P(Card_1 = A\spadesuit) = 1/52$. Let's say that, instead of drawing the Ace of Spades on the first draw, we drew the 3 of Hearts instead. What, then, is the probability that the second card is the Ace of Spades, given that we have already removed the 3 of Hearts from the deck? Clearly: $P(Card_2 = A\spadesuit)|Card_1 = 3\heartsuit) = 1/51$.

Joint and conditional probabilities are related to each-other:

$$P(X, Y) = P(X|Y) \times P(Y) = P(Y|X) \times P(X) \tag{110}$$

And by extension:

$$P(X|Y) = \frac{P(X, Y)}{P(Y)} \tag{111}$$

From these relationships we can derive Bayes Rule:

$$P(X|Y) = \frac{P(Y|X) \times P(X)}{P(Y)} \tag{112}$$

Bayes Rule is a profound fact about probabilities that deserves (and has received) far more attention than we can give it here.

*Independent Events*

If two variable events $X$ and $Y$ are independent, then the outcome of one event has no effect on, or discloses no information about, the state of the other. Two coin flips from the same coin are generally assumed to be independent: flipping heads on the first flip has no effect on the probability of getting heads on the next flip.

This is formalized by the conditional probability:

$$P(X|Y) = P(X) \iff X \perp Y \tag{113}$$

The probability of a particular event $X$ is unaffected by the outcome of $Y$.

Assuming independence, the joint probability of two events is equal to the product of the marginals:

$$P(X,Y) = P(X) \times P(Y) \iff X \perp Y \tag{114}$$

**Expected Values**

For some probability $P(X)$ and an associated value function $f : \mathcal{X} \mapsto \mathbb{R}$, then the *expected value* of $X$ is given by:

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} P(x) \times f(x) \tag{115}$$

For example, consider a fair coin $C$, with a support set $\{H, T\}$, and that the probability of both is $1/2$. We will also say that if $C$ flips $H$, then you, the player, get \$5.00 and if it flips $T$, you loose \$3.00. The long-term expected value of this game can be quantified with:

$$\mathbb{E}[C] = \frac{1}{2} \times \$5.00 + \frac{1}{2} \times -\$3.00 = \$1.00 \tag{116}$$

So, on average, over many trials, you would make $1.00 each turn. On the other hand, if the coin were weighted so that $T$ came up 80% of the time, the outcome would be quite different:

$$\mathbb{E}[C] = \frac{1}{5} \times \$5.00 + \frac{4}{5} \times -\$3.00 = -\$1.40 \tag{117}$$

So in this case you would *lose* money, on average, each term.

**Common Logical Gates**

*Logical AND*

| $P(x_1, x_2, y)$ | $X_1$ | $\wedge$ $X_2$ | $=$ $Y$ |
|---|---|---|---|
| 1/4 | 0 | 0 | 0 |
| 1/4 | 0 | 1 | 0 |
| 1/4 | 1 | 0 | 0 |
| 1/4 | 1 | 1 | 1 |

*Logical OR*

| $P(x_1, x_2, y)$ | $X_1$ | $\vee$ $X_2$ | $=$ $Y$ |
|---|---|---|---|
| 1/4 | 0 | 0 | 0 |
| 1/4 | 0 | 1 | 1 |
| 1/4 | 1 | 0 | 1 |
| 1/4 | 1 | 1 | 1 |

| $P(x_1, x_2, y)$ | $X_1$ | $\oplus$ $X_2$ | $=$ $Y$ |
|---|---|---|---|
| 1/4 | 0 | 0 | 0 |
| 1/4 | 0 | 1 | 1 |
| 1/4 | 1 | 0 | 1 |
| 1/4 | 1 | 1 | 0 |

*Logical Exclusive-OR (XOR)*